HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

BACHELORARBEIT

Frau
Tina Giersch

Charakterisierung biologisch relevanter Texte anhand von Wissenslandkarten

Mittweida, 2011

Fakultät Mathematik/
Naturwissenschaften/Informatik

BACHELORARBEIT

Charakterisierung biologisch relevanter Texte anhand von Wissenslandkarten

Autor:
Frau
Tina Giersch

Studiengang:
Biotechnologie/Bioinformatik

Seminargruppe:
Bi08w1-B

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Dipl.-Informatiker (FH) Daniel Stockmann

Einreichung:
Mittweida, 22.08.2011

Verteidigung/Bewertung:
Mittweida, 2011

Faculty Mathematica/
Natural Sciences /Computer Science

BACHELORTHESIS

characterising of biological relevant texts on the basis of KnowledgeMaps

author:
Ms.
Tina Giersch

course of studies:
biotechnology/bioinformatics

seminar group:
Bi08w1-B

first examiner:
Prof. Dr. rer. nat. Dirk Labudde

second examiner:
Dipl.-Informatiker (FH) Daniel Stockmann

submission:
Mittweida, 22.08.2011

defence/evaluation:
Mittweida, 2011

Bibliografische Beschreibung:

Giersch, Tina:

Charakterisierung biologisch relevanter Texte anhand von Wissenslandkarten. - 2011. - 5, 38, 4 S. Mittweida, Hochschule Mittweida, Fakultät Mathematik/Naturwissenschaften/ Informatik, Bachelorarbeit, 2011

Referat:

Die vorliegende Arbeit beschäftigt sich mit der Textanalyse biologisch relevanter Texte. Es geht darum Informationen und relevante Relationen aus diesen Texten zu extrahieren. Dies geschieht zunächst manuell und im Anschluss mit dem Programm antconc 3.2.1w maschinell. Anschließend wird mit dem Themengebiet *Wissenslandkarten* eine mögliche Form der Darstellung solcher Informationen und Zusammenhänge vorgestellt. Die nötigen Arbeitsschritte für die Erstellung einer solchen Wissenslandkarte werden näher beleuchtet, sowie deren Aufgaben, Ziele und Anwendungsgebiete dargestellt. Es wird auf die verschiedenen Arten von Wissenslandkarten erläutert und auf Vor- und Nachteile eingegnagen, die sich aus einer solchen Form der grafischen Darstellung ergeben. Desweiteren werden verschiedene Softwaretools vorgestellt, die die Erstellung einer Wissenslandkarte unterstützen können.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ABL	Acute Basophilic Leukemia
APA	Alternative Polyadenylation
ASCII	American Standard Code for Information
GEA	Gene Expression Analysis
CellGE	Cell Gene Expression
CGE	Cancer Gene Expression
HTML	Hypertext Markup Language
PCR	Polymerase Chain Reaction
TXT	TeXT
XML	eXtensible Markup Language

1 Einleitung

1.1 Zielsetzung

Ziel meiner Arbeit ist es, aus biologisch relevanten Texten relevante Informationen und Zusammenhänge zu extrahieren. Während in Kapitel zwei die Relationen zwischen diesen Texten im Anschluss an die Analyse manuell dargestellt wurden, möchte ich mich anschließend damit beschäftigen, wie dies maschinell geschehen kann und die Informationen als so genannte Wissenslandkarte dargestellt werden können.

Aufgrund der rasant ansteigenden uns zur Verfügung stehenden Daten möchte ich somit eine Möglichkeit aufzeigen wie Textsammlungen übersichtlich strukturiert werden können. Wissenslandkarten schaffen grafisch-visuell einen guten Überblick über große Datenmengen und erleichtern das Auffinden relevanter Informationen.

1.2 Kapitelübersicht

Die vorliegende Arbeit setzt sich aus vier Kapiteln zusammen. Im **ersten Kapitel** werden kurz die Ziele der Arbeit dargelegt und der Aufbau der Arbeit erläutert.

In **Kapitel zwei** werden zunächst wichtige Begriffe definiert, sowie die Entstehung von Wissen und verschiedene Wissensarten dargestellt. Im Anschluss werden verschiedene Text Mining-Werkzeuge vorgestellt und mit diesen eine manuelle und maschinelle Analyse einer Textsammlung bestehend aus 30 biologischen Texten durchgeführt. Anschließend werden die Ergebnisse in einem manuell erstellten Venn-Diagramm dargestellt. Abschließend beinhaltet dieses Kapitel ein kurzes Zwischenfazit.

Das **dritte Kapitel** handelt von Wissenslandkarten. Zunächst wird der Begriff *Wissenslandkarte* abgegrenzt. Es wird beschrieben, welche Schritte für die Erstellung einer solchen Karte nötig sind. Außerdem werden die Aufgaben und Ziele, Arten, sowie Vor-und Nachteile von Wissenslandkarten genannt und beschrieben. Es werden einige Softwaretools vorgestellt, die eine Erstellung von Wissenslandkarten unterstützen. Desweiteren wird die Erstellung und Ansicht einer Wissenslandkarte am Beispiel der KnowTech 2004 verdeutlicht.

Kapitel vier beinhaltet eine Zusammenfassung der beiden vorangegangenen Kapitel und liefert einen Ausblick.

2 Grundlagen

Möchte man sich mit der Charakterisierung von Texten beschäftigen, ist es zunächst einmal nötig, eine Definition für den Begriff *Text* festzulegen. Desweiteren soll die Entstehung von Wissen geklärt und verschiedene Wissensarten dargestellt werden.

2.1 Definition „Text“

Text repräsentiert Wissen und stellt eine Menge unstrukturierter Daten dar. „Mit Hilfe von Text Mining-Werkzeugen können aus digital vorliegenden Texten neue und relevante sachliche und inhaltliche Zusammenhänge extrahiert und strukturiert werden.“

[Heyer et al. 2006] S.1

2.2 Wissensentstehung

Laut [Ott 2003] umfasst die Entstehung von Wissen die folgenden Schritte: Daten, Informationen, Wissen und Aktion.

Daten stellen dabei das Rohmaterial für die Informationen dar, aus ihnen kann jedoch noch kein konkreter Wert abgeleitet werden. Um aus Daten *Informationen* gewinnen zu können, müssen die Daten beobachtet, gemessen, geordnet und strukturiert werden. Diese Informationen werden in einen Kontext von Relevanz für ein konkretes System eingebunden. Erst das Ergebnis aus der Verarbeitung und Interpretation von Informationen durch Intelligenz, Bewusstsein und Lernen bezeichnet man als *Wissen*. Mit Hilfe von Wissen lassen sich aus den Informationen Entscheidungen abzuleiten oder sich Erfahrungen zu Nutze zu machen und aus diesen zu lernen. Wissen stellt somit die Vorstufe für den Schritt der *Aktion* dar. [Ott 2003]

2.3 Wissensarten

Wissen kann in den unterschiedlichsten Formen auftauchen. Nach [Ott 2003] lassen sich konkret die folgenden Kategorien von Wissen definieren:

- *Verborgenes Wissen*: Wissen dass nicht mit Worten ausgedrückt werden kann
- *Verinnerlichtes Wissen*: Erfahrungen mit physischer Präsenz
- *Kodiertes Wissen*: Wissen, das trotz Verlust eines Mitarbeiters vorhanden bleibt
- *Konzeptionelles Wissen*: Kognitive Fähigkeit, Basisannahmen um übergeordnete Muster erkennen und überdenken zu können
- *Sozial konstruiertes Wissen*: Geteiltes Wissen wird aus verschiedenen Sprachsystemen (Organisations-)Kulturen, (Arbeits-) Gruppen etc. entwickelt
- *Ereigniswissen*: Wissen über Ereignisse und Trends
- *Prozesswissen*: Wissen zu Abläufen und Zusammenhängen

2.4 Ausgangspunkt

Meine Arbeit begann damit, geeignete Programme für die Textanalyse auszuwählen. Deren Evaluierung erfolgte zunächst an beliebigen Beispieltexten. Ich entschied mich dafür, im weiteren Verlauf meiner Arbeit mit dem Programm *antconc3.2.1w* [Anthony, 2011] zu arbeiten. Dieses Programm dient der Textanalyse eigener Korpora. Es bietet eine gute Benutzeroberfläche, vereint viele verschiedene Funktionen und erlaubt somit eine vielseitige Anwendung.

Der nächste Schritt bestand darin, eine geeignete Textsammlung für die weiteren Untersuchungen zu gewinnen. Hierfür erstellte ich einen Datensatz mit 30 Abstracts aus der Datenbank *PubMed*. [NCBI 2011] Alle 30 Abstracts gehören zu dem Überbegriff *gene expression*. Als Beispieltexte sollten mir jeweils die ersten 10 Ergebnisse zu den Suchanfragen *cancer gene expression* (CGE), *cell gene expression* (CellGE) und *gene expression analysis* (GEA) dienen.

2.5 Charakterisierung der Texte

Zur Charakterisierung der Texte dienten mir die Tools *Wortliste*, *N-Gramm* und *Kollokation*.

Wortliste

Bei der *Wortliste* handelt es sich um eine Liste aller Worttypen eines Textes, einschließlich deren Frequenz. Worttypen sind definiert als die Anzahl unterschiedlicher Wörter in einem Text. Betrachtet man die Anzahl aller vorhandenen Wörter, spricht man von Worttoken.

Beispiel:

Acute basophilic leukemia (ABL) is a rare subtype of acute leukemia with clinical features and symptoms related to hyperhistaminemia due to excessive growth of basophils.

Dieser Beispielsatz beinhaltet 25 Worttoken, jedoch nur 21 Worttypen. Die sich aus diesem Satz ergebende Wortliste sieht wie folgt aus (Tabelle 1):

Tabelle 1: Wortliste zum Beispielsatz

Frequenz	Wort
2	acute
2	leukemia
2	of
2	to
1	a
1	abl
1	and
1	basophilic
1	basophils
1	clinical
1	due
1	excessive
1	features
1	growth
1	hyperhistaminemia
1	is
1	rare
1	related
1	subtype
1	symptoms
1	with

Die Reihenfolge der Wörter in der Liste ergibt sich aus deren Frequenz. Die Worttypen mit der höchsten Frequenz stehen am Anfang der Liste, Wörter mit der niedrigsten Frequenz am Schluss. Haben mehrere Worttypen dieselbe Frequenz, werden sie alphabetisch geordnet.

Im Beispiel wurden alle Buchstaben als Kleinbuchstaben behandelt, das Programm entfernt automatisch alle Satzzeichen.

N-Gramm

Bei N-Grammen handelt es sich um eine Wortfolge mit *n* Wörtern. Wenn *n = 2* spricht man von Bigrammen, falls *n = 3* von Trigrammen usw. Meist ist es ausreichend Bi- und Trigramme eines Textes zu betrachten. N-Gramme höherer Stufe vergrößern nur den Rechen- und Speicheraufwand, bringen jedoch kaum noch eine Verbesserung in der Genauigkeit.

Beispiel:

We used this strategy to explore the profiling of alternative polyadenylation (APA) sites in two human breast cancer cell lines, MCF7 and MB231, and one cultured mammary epithelial cell line, MCF10A.

Betrachtet man Bigramme, wird ein Fenster der Größe zwei festgelegt. Dieses Fenster rückt, wie im Beispielsatz angedeutet, stets um ein Wort weiter. Die folgende Tabelle (Tabelle 2) zeigt die Bigramm-Liste zum Beispielsatz. Auch hier werden die Ergebnisse nach ihrer Frequenz und alphabetisch geordnet. Rot umrahmt wurde der Bereich, in dem die, im Beispielsatz markierten, Bigramme zu finden sind.

Tabelle 2: Bigramme im Beispielsatz

Frequenz	Bigramm
1	and MB
1	and one
1	APA sites
1	breast cancer
1	cancer cell
1	cell line
1	cell lines
1	cultured mammary
1	epithelial cell
1	explore the
1	human breast
1	in two
1	line MCF
1	lines MCF
1	mammary epithelial
1	MB and
1	MCF A
1	MCF and
1	of APA
1	one cultured
1	profiling of
1	sites in
1	strategy to
1	the profiling
1	this strategy
1	to explore
1	two human
1	used this
1	We used

Kollokation

Als Kollokation bezeichnet man Ausdrücke bestehend aus mehreren Wörtern. Um auch getrennte Wortkombinationen als solche zu erkennen, wird ein Fenster definierter Größe auf beiden Seiten des Wortes, zu dem eine Kollokation gesucht wird, festgelegt.

Beispiel:

…the thyroid hormone response element-bound TRα1PV than to TRβ1PV in the promoter of the CCAAT/ enhancer-binding protein α ***gene*** *to repress its* ***expression*** *in the liver of Thra1(PV) mice, …*

Die Fenstergröße zum Suchwort *gene* beträgt in diesem Fall je vier Worte nach links und nach rechts. Die Wortkombination *gene expression* wird trotz der eingeschobenen Worte „to repress its“ als solche gefunden.

2.5.1 Manuelle Analyse

Um später überprüfen zu können ob das verwendete Programm *antconc3.2.1w* funktioniert, wurden die 30 Texte zunächst per Hand analysiert. Das heißt alle Texte wurden gelesen und wichtige Wörter, Wortgruppen, Kollokationen, sowie deren Frequenzen wurden extrahiert. Die Ergebnisse sollen hier auszugsweise an dem Abstract „Prognostic *gene-expression* signature of *carcinoma*-associated *fibroblasts* in non-small cell *lung cancer*“ (siehe Anlagen, Teil 1) dargestellt werden. Dieser Abstract erschien als erstes Suchergebnis zu der Suchanfrage *cancer gene expression*.

Die Ergebnisse der manuellen Analyse wurden im Abstract kursiv hervorgehoben und werden in der folgenden Tabelle (Tabelle 3) dargestellt.

Tabelle 3: Ergebnisse der manuellen Textanalyse

Wort	**Frequenz**
cancer	6
carcinoma	3
fibroblasts	3
Bigramm	
gene expression	5
lung cancer	2
cancer development	1
cancer institute	1
cancer research	1
expressed genes	1
lung carcinoma	1
Trigramm	
cell lung cancer	1
Kollokation	
genes also were differentially ***expressed***	1

2.5.2 Maschinelle Analyse

Für die maschinelle Analyse des Datensatzes mit dem Programm *antconc3.2.1w* war zunächst eine Konvertierung der Texte erforderlich. Die Dateien müssen im TXT-, HTML- oder XML-Format vorliegen, alle Buchstaben wurden in Kleinbuchstaben umgewandelt und Zeilenumbrüche in Leerzeichen konvertiert.

Auch für die maschinelle Analyse werden die Ergebnisse nur auszugsweise an dem in der Anlage aufgeführten Abstract *„Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer“* dargestellt.

Für eine bessere und übersichtlichere Darstellung der Ergebnisse wurden in allen drei Untersuchungsschritten (Wortliste, N-Gramm und Kollokation) irrelevante Wortgruppen und (Füll-) Wörter wie bspw. *of*, *the*, *and*, *in* etc. manuell aus den Ergebnis-Listen heraus gelöscht.

Die Ergebnisse der Wortliste können Tabelle 4 entnommen werden. Das Resultat der N-Gramm-Untersuchung wird in Tabelle 5 dargestellt.

Tabelle 4: Wortliste der maschinellen Analyse

Rank	Frequenz	Wort
1	6	cancer
6	5	expression
7	5	gene
9	4	cell
10	4	genes
11	4	lung
13	4	nsclc
17	3	carcinoma
18	3	fibroblasts
21	3	prognostic
29	2	expressed
34	2	lines
35	2	microarray
36	2	nf
37	2	nfs
40	2	primary
41	2	protein
42	2	proteins
44	2	signaling
45	2	signature
46	2	small
47	2	stroma
48	2	tgf

Rank	Frequenz	Wort
50	2	tumor
57	1	analyses
58	1	analysis
59	1	annotation
60	1	apr
73	1	clinical
78	1	cultures
84	1	encoding
90	1	extracellular
92	1	family
97	1	functional
106	1	induced
107	1	influences
108	1	interaction
117	1	metastasis
118	1	microdissected
119	1	microenvironment
121	1	multiple
130	1	pathway
131	1	pathways
132	1	patients
134	1	pmid
169	1	tumorigenicity

Tabelle 5: Maschinelle Untersuchung auf N-Gramme

Rank	Frequenz	Bigramm
1	5	gene expression
6	2	cell lines
7	2	cell lung
12	2	lung cancer
18	2	prognostic gene
59	1	cancer cell
60	1	cancer development
61	1	cancer institute
63	1	cancer research
64	1	cancer stroma
88	1	expressed genes
90	1	expression analysis
91	1	extracellular proteins
108	1	genes encoding
109	1	genes probe
129	1	influences cancer
137	1	laser capture
140	1	lines microarray
141	1	lung carcinoma
152	1	microarray gene
161	1	nf cell
168	1	nsclc microarray
169	1	nsclc patients
183	1	primary cultures
184	1	primary tumor
192	1	prominent involvement
193	1	protein interaction
194	1	protein protein
195	1	proteins regulated
248	1	tumor microenvironment
249	1	tumor stroma
		Trigramm
2	2	carcinoma associated fibroblasts

Bei der Untersuchung auf Kollokationen wurde im hier aufgeführten Beispiel nach Kollokationen zu den Worten *cancer* und *gene* (Search Term) gesucht. Die Fensterlänge betrug dabei je vier Worte nach links und nach rechts (4L-4R). In der Ergebnis-Liste wird die gesamte Frequenz dargestellt, es wird jedoch auch angezeigt, wie oft die Kollokation links (Freq-L) oder rechts (Freq-R) vom Suchwort zu finden ist. Die Ergebnisse sind in Tabelle 6 dargestellt.

Tabelle 6: Kollokationen 4L-4R maschinelle Analyse

Search Term	Rank	Freq	Freq-L	Freq-R	Collocate
cancer	6	2	2	0	lung
	7	2	0	2	gene
	9	2	1	1	cell
	11	1	1	0	tumorigenicity
	12	1	1	0	tumor
gene	2	5	0	5	expression
	7	2	2	0	cancer

2.6 Visualisierung der Ergebnisse

Für eine Visualisierung meiner Ergebnisse, musste zunächst eine Einteilung der Texte vorgenommen werden. Hierfür definierte ich für mich relevante Schlüsselwörter und erstellte mit diesen eine geordnete Liste für die Klassifizierung. Die Ergebnisse können der Tabelle 8 (siehe Anlagen, Teil 2) entnommen werden. Rotmarkierte Wörter bedeuten, dieses Wort dient in dieser Textklasse (*cancer gene expression* CGE, *cell gene expression* CellGE oder *gene expression analysis* GEA) als Schlüsselwort; blaumarkiert sind Wörter, die in einer anderen Textklasse als Schlüsselwort dienen. Kreuze bedeuten, das Wort kommt in dem entsprechenden Text vor.

Für eine bessere Übersicht wurden die Ergebnisse als Histogramm dargestellt (Abbildung 1). Es wird die Anzahl der Schlüsselwörter in den drei Textklassen abgebildet.

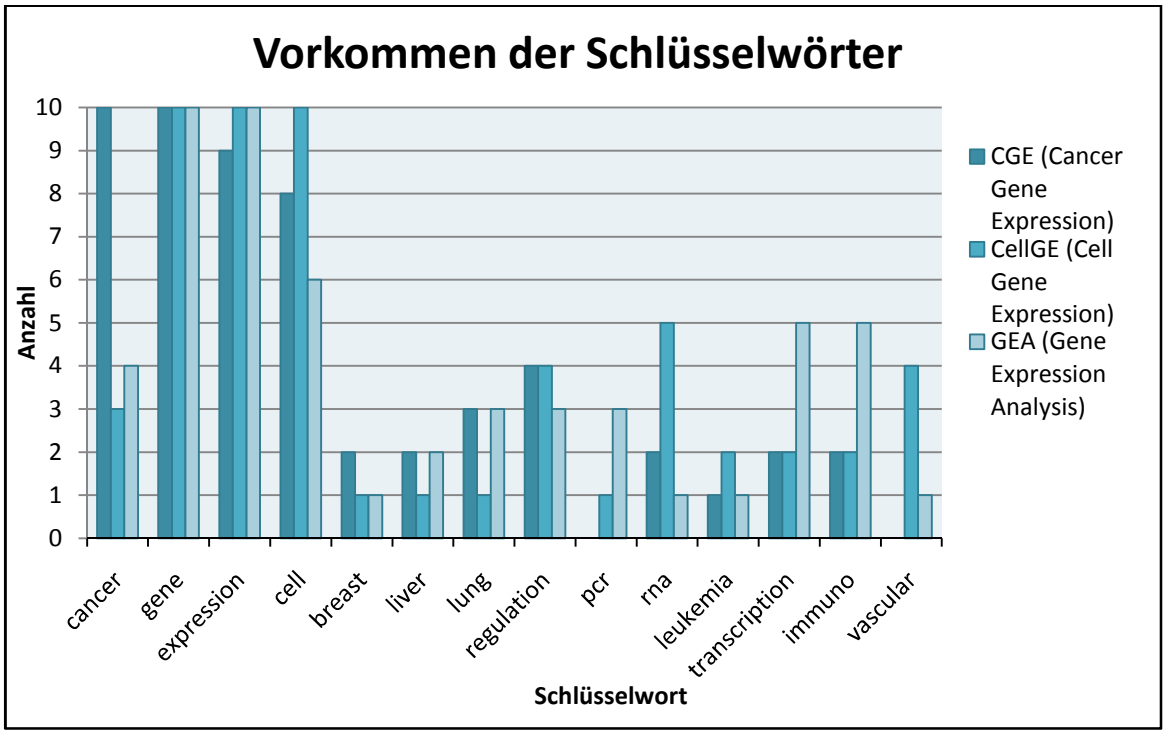

Abbildung 1: Vorkommen der Schlüsselwörter

Unter Einbezug aller vorangegangenen Ergebnisse, habe ich anschließend das folgende Venn-Diagramm erstellt (Abbildung 2).

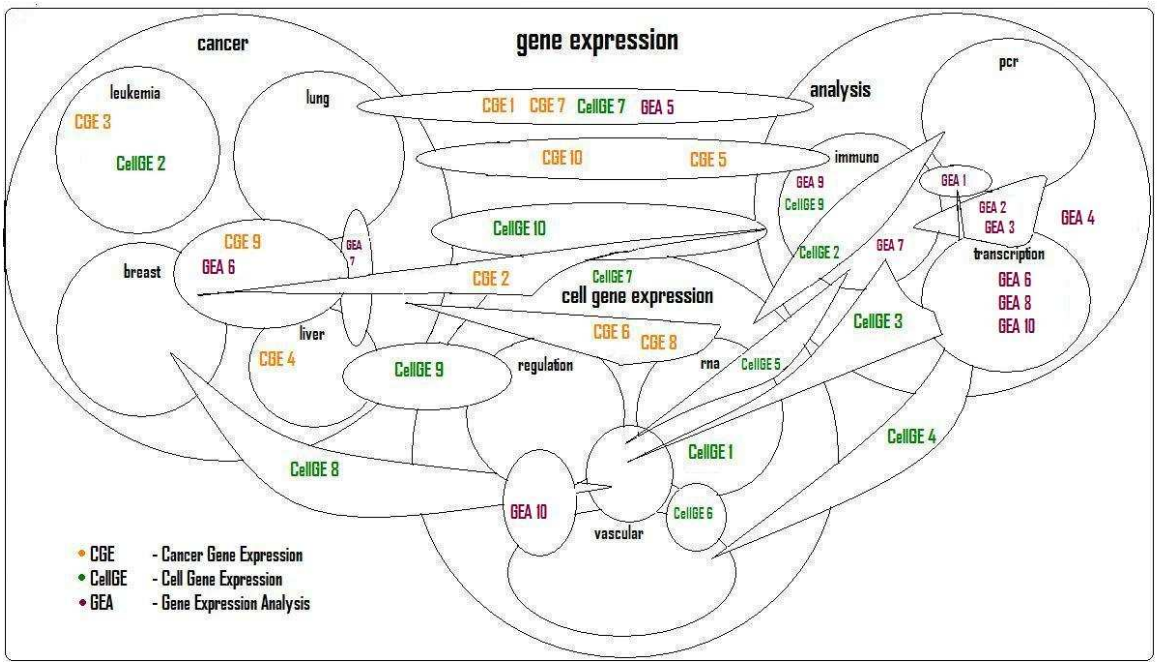

Abbildung 2: Venn-Diagramm

Aufgrund der Schlüsselwort-Suche wäre zu vermuten gewesen, dass lediglich Relationen zwischen Texten innerhalb einer Textklasse (*cancer gene expression*, *cell gene expression* oder *gene expression analysis*) existieren. Die Abbildung zeigt jedoch, dass auch zwischen Texten aus verschiedenen Textklassen Verbindungen bestehen.

Beispielsweise weist der Text GEA 6 die Worte *lung*, *liver* und *breast* auf und ist somit näher mit dem Text CGE 9 verwandt, als mit anderen Texten seiner Textklasse. Ebenfalls ist zum Beispiel der Text CGE 4 näher mit dem Text CGE 9 verwandt, als mit anderen cancer gene expression Texten, da es in diesen beiden Texten um Leberkrebs geht. Die Texte CGE 5 und CGE 10 beinhalten unter anderem jeweils das Schlüsselwort *immuno* und stehen somit den Texten GEA 9 und CellGE 9 nahe usw.

2.7 Zwischenergebnis

Als Abschluss dieses Kapitels lässt sich festhalten, dass die manuelle Analyse das Ergebnis der maschinellen Analyse bestätigen konnte. Aus den Texten konnten mit beiden Methoden gleichermaßen Informationen und Zusammenhänge extrahiert werden.

Wie in der Abbildung 2 verdeutlicht wird, deckte die Evaluierung andere Relationen zwischen den Texten auf, als man anfangs nach der Suche mit den Such-Terms *cancer gene expression*, *cell gene expression* und *gene expression analysis* erwartet hätte.

3 Wissenslandkarten

Während in Kapitel zwei ein Großteil der Auswertung, so wie die abschließende grafische Darstellung der Ergebnisse manuell geschah, möchte ich mich in diesem Kapitel mit einer Darstellungsform beschäftigen, die es ermöglicht Texte automatisch auszuwerten und anschließend eine Übersicht von relevanten Schlüsselwörtern und deren semantischer Vernetzung zu erstellen.

3.1 Begriffsdefinition „Wissenslandkarte“

Als Wissenslandkarte bezeichnet man ein semantisch strukturiertes Wortnetz, mit dem in Form eines Grafen eine Übersicht wesentlicher Schlüsselwörter einer Textsammlung und deren inhaltliche Vernetzung dargestellt werden können. [Heyer et al. 2006]

Synonyme sind Wissenskarte, KnowledgeMap oder Wissensnetz.

Um eine Wissenslandkarte berechnen zu können, müssen die Texte mit verschiedenen Text Mining-Verfahren bearbeitet werden. Diese beinhalten eine Differenz- und Kookkurenzanalyse, wesentliche Themenschwerpunkte und deren Schlüsselbegriffe müssen hervorgehoben werden, Autoren werden ihren Themenschwerpunkten, sowie ihren Firmen zugeordnet bis schließlich die Vernetzung von Themen, Autoren und Firmen in Form eines semantisch strukturierten Themennetzes stattfinden kann.

Damit diese Textanalyse stattfinden kann, sind zunächst jedoch auch einige umfangreiche textuelle Vorbereitungsschritten, wie die Konvertierung der Texte, eine Segmentierung der Texte in Sätze und Wortformen, sowie eine Grundformreduktion nötig.

[Heyer et al. 2006]

3.2 Erstellen einer Wissenslandkarte

Um eine Wissenslandkarte berechnen zu können, müssen die Texte mit verschiedenen Text Mining-Verfahren bearbeitet werden. Diese beinhalten eine Differenz- und Kookkurenzanalyse, wesentliche Themenschwerpunkte und deren Schlüsselbegriffe müssen hervorgehoben werden, Autoren werden ihren Themenschwerpunkten, sowie ihren Firmen zugeordnet bis schließlich die Vernetzung von Themen, Autoren und Firmen in Form eines semantisch strukturierten Themennetzes stattfinden kann.

3.2.1 Textuelle Vorverarbeitung

Damit diese Textanalyse stattfinden kann, sind zunächst einige umfangreiche textuelle Vorbereitungsschritten, wie die Konvertierung der Texte, eine Segmentierung der Texte in Sätze und Wortformen, sowie eine Grundformreduktion nötig.

Alle Texte der vorliegenden Textsammlung sollten zunächst einheitlich in das ASCII-Format konvertiert werden. Im Anschluss müssen aus jedem Text die Titel, Autoren und deren Organisationen extrahiert werden.

3.2.1.1 Segmentierung von Text in Sätze und Wortformen

Die Zerlegung von Texten in Sätze und Wortformen ist für den Menschen problemlos zu lösen und auch maschinell meistens korrekt möglich. Jedoch gibt es auch einige Ausnahmen, bei denen die allgemeinen Verfahren keine Anwendung finden können und zusätzliche Regeln oder Ausnahmelisten erstellt werden müssen.

Regeln für die Satzsegmentierung

Zwischen einem Satzanfang und einem Satzende befindet sich stets eine Trennstelle. Daher konnten folgende Regeln für den Satzanfang bzw. für das Satzende aufgestellt werden.

Regeln für den Satzanfang:

- Nach einer Überschrift beginnt ein neuer Satz.
- Sätze beginnen nie mit Kleinbuchstaben.
- Am Anfang eines Abschnittes beginnt ein neuer Satz.
- Großgeschriebene Artikel sind Indizien für einen neuen Satz.
- Falls kein neuer Absatz beginnt, steht vor dem neuen Satz ein Satzendzeichen.

Regeln für das Satzende:

- Sätze enden mit Satzendzeichen (Punkt, Fragezeichen oder Ausrufezeichen).
- Nach dem Satzendzeichen muss ein white space (Leerzeichen, Tabulator, Zeilenumbruch) stehen.
- Vor einer Überschrift endet ein Satz.
- Am Ende eines Absatzes endet ein Satz.
- Überschriften sollten wie Sätze behandelt werden.

- Beginnt das folgende Wort nach einem Satzendzeichen mit einem Kleinbuchstabe, handelt es sich nicht um ein Satzende.

[Heyer et al. 2006] S. 63

Schwierige Fälle, bei denen diese Regeln jedoch versagen würden, sind beispielsweise:

Er trägt den Titel Dr. rer. nat. Schulze.

Um Abkürzungen als solche und nicht fälschlicherweise als Satzende zu interpretieren, kann einfach eine Liste mit Abkürzungen erstellt werden. Von Bedeutung sind dabei jedoch nur Abkürzungen, die mit einem Punkt enden, da es sonst zu keiner Verwechselung kommen kann. Eine solche Liste kann dann wie folgt aussehen (siehe Tabelle 7).

Tabelle 7: Häufige deutsche Abkürzungen [Heyer et al. 2006] S. 65

Rang	**Abkürzung**
1	Mio.
2	Tel.
3	Dr.
4	bzw.
5	Nr.
6	Co.
7	Mill.
8	u.
9	Sa.
10	a.
11	Prof.
12	u.a.
13	Str.
14	Kl.
15	ca.
16	e.V.
17	z.B.
18	z. B.
19	Bekl.
20	St.

Mit einer solchen Liste oder Mustern wie zum Beispiel **str.* oder **ges.* für Abkürzungen wie *Humboldtstr.* oder *Handelsges.* lassen sich die Schwierigkeiten mit den Abkürzungen recht gut beheben.

Probleme bleiben jedoch bei einem Punkt nach einer Zahl wie im Beispiel *Morgen ist Freitag der 13. Juli.*

Schwierigkeiten gibt es auch bei der wörtlichen Rede:

„Ich kann es hören! Es kommt immer näher“, rief er entsetzt. [Heyer et al. 2006] S. 64

In diesem Beispiel würde es laut den oben aufgestellten Regeln zu folgender Zerlegung kommen:

„Ich kann es hören!

Es kommt näher“, rief er entsetzt.

Nun treten in beiden Sätzen isolierte An- bzw. Ausführungszeichen auf, so dass dieses Ergebnis nicht zufriedenstellend ist. Für den ersten Satz kann die Lösung lauten: Ist das erste oder letzte Zeichen eines Satzes ein isoliertes An- oder Ausführungszeichen, so entferne es. Für ein solches isoliertes An- oder Ausführungszeichen mitten im Satz gibt es jedoch keine vergleichbar einfache Lösung.

Ein weiteres Problem kann sich ergeben, wenn Listen oder Programmcode nicht als solche erkannt werden, sodass extrem lange Sätze entstehen. Die Lösung kann in diesem Fall einfach sein, dass man eine maximale Satzlänge von beispielsweise 200 Wörtern festlegt und längere Objekte ignoriert werden. Dabei kann es jedoch auch passieren, dass korrekte Sätze mit einer Länge oberhalb dieser Grenze nicht erfasst werden. Eine solche Schranke ist also mit Bedacht festzulegen und es ist zu bedenken, ob ein möglicher Verlust solcher Sätze zu akzeptieren ist.

Segmentierung in Wortformen

Wortformen werden sowohl im Deutschen als auch im Englischen durch einen white space getrennt. Da häufig von folgender Definition für Wortformen ausgegangen wird, scheint die Trennung am white space richtig zu sein.

„Wortformen sind die in einem grammatisch und orthographisch korrekten Text stehenden Zeichenketten bestehend aus Buchstaben und Bindestrichen, die im Text durch white space oder Satzzeichen getrennt sind.“ [Heyer et al. 2006] S. 66

Dass dies jedoch nicht immer zutreffend ist, verdeutlichen folgende Beispiele:

Alternative Energiequellen haben Vor- und Nachteile.

Der SAT 1-Moderator berichtete vor Ort.

Im ersten Beispiel ist *Vor-* kein Wort und erhält seinen Sinn erst in der Fügung *Vor- und Nachteile.*

Im zweiten Beispiel würde man durch die Trennung am white space *1-Moderator* erhalten obwohl nur *SAT 1-Moderator* eine sinnvolle Wortform ergibt.

Weitere Fehler in den Wortformen können durch Rechtschreibfehler oder Worttrennungen, die ehemals am Zeilenende standen, entstehen.

Desweiteren ergeben sich Schwierigkeiten aus wortähnlichen Objekten, wie zum Beispiel *Internetadressen*, *Dateinamen*, *chemische Formeln*, *mathematische Ausdrücke*, etc.

Hierfür ist es sinnvoll zu prüfen, ob die nach der Wortsegmentierung erhaltenen Wortformen aus einer Menge erlaubter Zeichen bestehen. Im Deutschen beispielsweise sind diese erlaubten Zeichen die Groß- und Kleinbuchstaben a bis z, ß, die Umlaute ä, ö, ü sowie Ä, Ö, Ü und der Bindestrich. Sinnvoll ist es auch das é, Ziffern in Eigennamen und Großbuchstaben folgend auf Kleinbuchstaben hinzuzunehmen, so dass Wortformen wie *Café*, *Audi A4* oder *pH-Wert* nicht zurückgewiesen werden. Verboten werden sollten jedoch andere Satzzeichen als der Bindestrich, somit können beispielsweise Dateinamen mit einem Punkt in ihrem Inneren ausgeschlossen werden.

3.2.1.2 Musteranalyse

Die Musteranalyse beinhaltet Schritte wie die Grundformreduktion und Terminologieextraktion, auf welche ich im folgenden Kapitel näher eingehen möchte.

Zunächst gilt es jedoch einige grundlegende Begriffe zu klären:

Morpheme: Als Morpheme werden „die kleinsten bedeutungstragenden Buchstabenkombinationen“ [Heyer et al. 2006] S.320 bezeichnet. Von freien Morphemen spricht man bei der Grundform der Wortform.

Allomorphe: „Freie Morpheme, die in ihrer Gestalt verschieden, aber semantisch identisch sind, werden als Allomorphe bezeichnet.“ [Heyer et al. 2006] S.329

Derivation: „Derivation schafft neue Wortformen durch Anfügung (Affigierung) von Derivativen an Wortstämme. […] Derivationssuffixe ändern zudem meist die Wortart des Stamms.“ [Heyer et al. 2006] S.320

Grundformreduktion

Bei einer Grundformreduktion von Wortformen wird eine Neutralisierung der Flexion durchgeführt, das heißt die Vollformen werden wieder auf ihre Grundform zurückgeführt. So können verschiedene Vollformen die derselben Grundform angehören und als semantisch gleichwertig angesehen werden zu einem Konzept zusammengefasst werden und erleichtern somit später die Analyse des Textes.

Im Englischen ist diese Aufgabe beispielsweise relativ einfach zu bewältigen, da diese Sprache nur wenige Flexive, wie zum Beispiel *–s*, *-ed* und *–ing* besitzt.

Im Deutschen hingegen gibt es eine sehr große Menge verschiedener Flexive, desweiteren stellen häufig auftretende Allomorphe eine weitere Herausforderung dar. Beim Anhängen von Derivationssuffixen wird häufig auch der Wortstamm verändert. Somit reicht es nicht, diese bei der Grundformreduktion einfach wieder abzuschneiden. Es benötigt ein komplexes Regelwerk und am besten auch ein umfangreiches Lexikon.

[Heyer et al. 2006]

Terminologieextraktion

Fachtexte verfügen meist über ein domänenspezifisches Vokabular. So beinhalten beispielsweise naturwissenschaftliche oder medizinische Texte oft Wörter mit domänenspezifischen lateinischen oder griechischen Suffixen. Als Beispiele wären hier zu nennen *–itis* in der Medizin oder *–ase* in der Chemie. Wortformen mit solchen spezifischen Endungen können mittels der Differenzanalyse, welche im Kapitel 3.2.2 näher beschrieben wird, ganz einfach herausgefiltert werden und man erhält eine Liste mit Wortformen, bei denen es sich mit hoher Wahrscheinlichkeit um Fachausdrücke handelt.

3.2.2 Differenzanalyse

Ein nächster und wichtiger Schritt für die Erstellung einer Wissenslandkarte ist die **Beschlagwortung** jedes Textes. Somit können die Texte individuell charakterisiert werden, aber auch inhaltlich verwandte Beiträge miteinander in Zusammenhang gesetzt werden. Die Beschlagwortung erfolgt mittels Differenzanalyse.

Als Grundlage für die Differenzanalyse dienen zwei Textmengen: ein **Analyse-** und ein **Referenzkorpus**. Als Analysekorpus, bezeichnet man die zu untersuchende Textmenge, in unserem Fall also die Abstracts vom Biotechnologie-Tag in Dresden. Im Gegensatz dazu ist der Referenzkorpus ein allgemeinsprachlicher Textkorpus, der beispielsweise aus Zeitungsartikeln erstellt wird. Nun können dann die Auftretenswahrscheinlichkeiten einzelner Wortformen oder Wortkombinationen für beide Korpora berechnet und im Anschluss verglichen werden. Dabei kann in folgende vier Klassen unterschieden werden:

Klasse 1: Kommt eine Wortform nicht im Referenzkorpus vor, handelt es sich mit hoher Wahrscheinlichkeit um Fachausdrücke des Themengebietes über das im Analysekorpus berichtet wird.

Klasse 2: Ebenfalls um Fachterme handelt es sich mit einer gewissen Wahrscheinlichkeit bei Wortformen, die im Analysekorpus relativ häufiger vorkommen als im Referenzkorpus. Für eine Identifizierung dieser Wortformen muss ein Schwellenwert festgelegt werden.

Klasse 3: Wenn Wortformen sowohl im Analyse- als auch im Referenzkorpus mit etwa der gleichen Wahrscheinlichkeit vorkommen, handelt es sich wahrscheinlich um Stoppwörter (Artikel, Präpositionen, Konjunktionen) oder allgemeine Begriffe. Diese Wörter geben keinen Aufschluss über den Inhalt der Beiträge des Analysekorpus.

Klasse 4: Kommen Wortformen im Analysekorpus seltener vor, als im Referenzkorpus, handelt es sich im Allgemeinen nicht um Fachausdrücke aus dem Fachbereich der zu analysierenden Texte.

Für das Anwendungsgebiet der Beschlagwortung ist es in der Regel ausreichend die Klassen 1 und 2 zu betrachten.

Für die Ermittlung von Schlagwörtern ist es jedoch auch wichtig, dass neben der Übergewichtung der Frequenz dieser Wortformen im Analysekorpus gegenüber der Frequenz im Referenzkorpus auch noch weitere Kriterien erfüllt werden:

- **Relevanz**: Mindestfrequenz im Analysekorpus festlegen, da wichtige Begriffe oft wiederholt auftreten

- **Bekanntheit**: Da wichtige Begriffe meist auch bekannte Begriffe sind, sollten sie auch im Vergleichskorpus mit einer Mindestfrequenz vorkommen.

- **Grundformen**: Wörter treten im Text meist in flektierter Form auf, sollten bei der Analyse jedoch in ihrer Grundform angezeigt werden.

- **Beschränkung auf Substantive**: Dieses Kriterium erscheint sinnvoll, da es sich bei wichtigen Meldungen in der Regel um Personen oder Ereignisse handelt. Verben, Adjektive oder Adverbien sind meist unspezifisch und geben oft keine konkrete Auskunft.

3.3 Aufgaben und Ziele von Wissenslandkarten

Nachdem nun ein Überblick über die einzelnen Schritte zur Erstellung einer Wissenslandkarte gegeben wurde, kann man sich natürlich die Frage stellen: Wozu werden solche Wissenslandkarten erstellt und welchen Nutzen habe ich davon?

Mit Hilfe modernen Wissensmanagements lassen sich Wissensressourcen in Unternehmen oder Organisationen optimaler ausnutzen. Wissenslandkarten verweisen auf „explizites oder implizites Wissen, das in externen oder internen Dokumenten, Datenbanken oder in den Köpfen von Experten vorhanden ist“[Stampf 2006]. Nach von Krogh und Venzin [1995] umfasst das Wissensmanagement folgende Aufgaben:

- Erschließen von Wissen (Erfahrungen, Best Practices) für alle, die dieses Wissen im Rahmen ihrer organisatorischen Rolle benötigen
- Verfügbarmachen von Wissen am Ort und zur Zeit der Entscheidung
- Erleichtern des effektiven und effizienten Entwickelns von neuem Wissen
- Sicherstellen, dass jeder in der Organisation weiß, wo Wissen verfügbar ist, und
- Umsetzung dieser Kompetenzen in neue Produkte und Dienstleistungen.

[Biemann et al. 2004]

Nach [Heyer et al. 2006] haben Wissenslandkarten auch die folgenden Aufgaben:

- Wissenslandkarten erleichtern die Orientierung innerhalb großer Dokumentenkollektionen
- in Texten verborgene Strukturen werden übersichtlich und intuitiv erfassbar dargestellt

Eine traditionelle Form des Wissensaustausches stellen Messen und Fachtagungen dar. Dabei kann eine Wissenslandkarte sowohl für Anbieter als auch für Nachfrager von großem Nutzen sein. Der Anbieter kann so in Erfahrung bringen, in welchem Wettbewerbsumfeld sich sein Angebot befindet, welches seine Mitbewerber sind und wo sein Angebot im Gesamtzusammenhang einschlägiger Angebote platziert ist. Der Nachfrager kann sich einen Überblick über für ihn interessante Angebote machen, Zusammenhänge zwischen verschiedenen Anbietern erkennen und kann seine „Anfrage im Gesamtzusammenhang der verfügbaren Angebote einordnen und bewerten“ [Biemann et al. 2004]. Da sich dieses Prinzip auch auf andere Fälle übertragen und verallgemeinern lässt, stellen Wissenslandkarten ein breites und vielfältiges Anwendungsgebiet dar.

Wissenslandkarten sollen Transparenz über unternehmensinternes und/oder externes Wissen schaffen, um so Ressourcen effizienter nutzen zu können und den Zugriff auf benötigtes Wissen zu beschleunigen. Große Datenmengen und komplexe Zusammenhänge können grafisch-visuell schneller und exakter erfasst werden, als es verbal oder durch Zahlenwerte möglich ist [Wissenslandkarte 2008].

3.4 Arten von Wissenslandkarten

Generell können Wissenslandkarten in fünf verschiedene Arten unterteilt werden: Wissensträgerkarten, Wissensbestandskarten, Wissensanwendungskarten, Wissensstrukturkarten und Wissensentwicklungskarten.

Wissensträgerkarten

Wissensträgerkarten haben hauptsächlich das Ziel Wissensträger zu identifizieren. Wissensträger können Personen sein, die über relevante Kompetenzen und Fachwissen verfügen, es kann sich jedoch auch um Bücher, Projekte, Organisationseinheiten oder ähnliches handeln. Diese Wissensträger werden somit bekannt gemacht und zu ihrem entsprechenden Wissensgebiet zugeordnet, so dass sie leicht von den Nutzern der Karte gefunden werden können.

Das Ziel von Wissensträgerkarten ist es, allen Mitarbeitern eines Unternehmens Zugang zu den für sie relevanten Informationen zu verschaffen. Solche Wissenslandkarten sollten über Informationen verfügen, die über Abteilungsgrenzen und optimalerweise auch über Unternehmensgrenzen hinaus gehen. So könnte es beispielsweise Neueinsteigern, die sich noch kein Beziehungsnetz innerhalb der Firma aufbauen konnten, erleichtern relevante Experten ausfindig zu machen. Die Kommunikation zwischen Wissensträgern und Wissenssuchenden kann mit Hilfe von Wissensträgerkarten wesentlich verbessert werden. [Ott 2003]

Wissensträgerkarten stellen die klassische Art von Wissenslandkarten dar und werden oft durch andere im nachfolgenden aufgeführte Wissenslandkarten ergänzt.

Ein Beispiel für eine Wissensträgerkarte stellt die Abbildung 3 dar. Die Abbildung wird unterteilt in fünf die Kernkompetenzen: *Animation, Didaktik, Inhalt, Kommunikation & Koordination,* so wie *Grafik & Design*. Anhand dieser Karte wird ersichtlich welche Person über welches Fachwissen verfügt. Zusätzlich wird angezeigt in welchem Ort sie arbeiten.

Der Nachteil einer Wissensträgerkarte wie sie in Abbildung 3 dargestellt ist, kann jedoch sein, dass einige Informationen verloren gehen. Es ist möglich, dass sich Mitarbeiter in mehreren Bereichen als kompetent erweisen, sie werden jedoch meist nur einmal in der Karte aufgeführt.

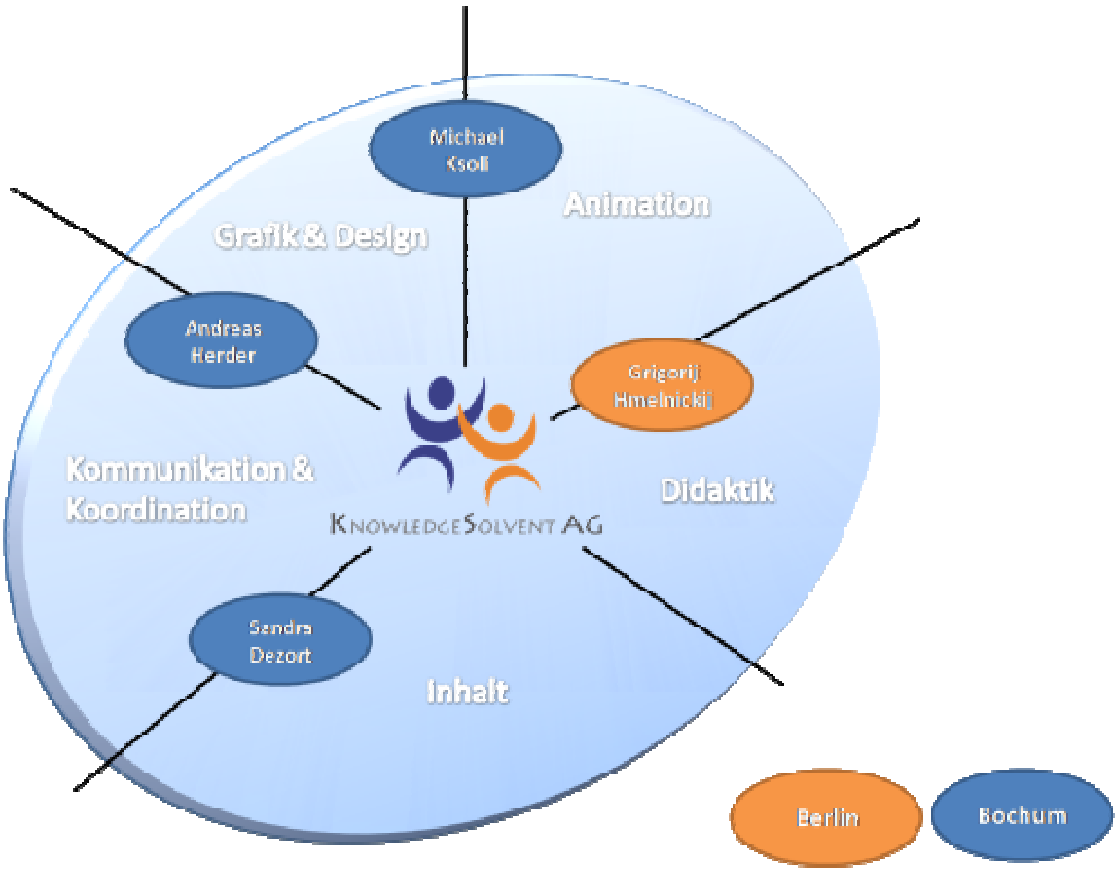

Abbildung 3: Beispiel einer Wissensträgerkarte [KS AG 2011]

Wissensbestandskarten

Wissensbestandskarten zeigen den Speicherort und den Umfang des Wissen an. Es werden mögliche Weiterverarbeitungsschritte und/oder quantitativ die Fähigkeiten der Mitarbeiter dargestellt. Desweiteren bekommt man mit Hilfe von Wissensbestandskarten auch Informationen über des Aggregatzustand des Wissens. Das heißt man erfährt nicht nur ob und wo das gesuchte Wissen vorhanden ist, sondern auch in welcher Form dieses Wissen abgelegt ist. Dies kann beispielsweise in einem Rechenzentrum, auf Papier oder auch im Gedächtnis eines Kollegen geschehen sein.

Auch diese Form der Wissenslandkarte hat das Ziel vorhandene Wissensbestände sichtbar zu machen und den Zugriff auf dieses Wissen zu erleichtern.

Als Veranschaulichung soll die Abbildung 4 dienen. Es wird deutlich, welcher Mitarbeiter in den Bereichen *Inhalt, Didaktik, Grafik & Design, Animation und Koordination & Kommunikation* in welchem Umfang über welches Wissen verfügt.

Vergleicht man nun diese Karte mit der Abbildung 3, wird auch hier noch einmal der Nachteil der Wissensträgerkarte deutlich. So wurde beispielsweise Andreas Herder in Abbildung 3 lediglich den Kompetenzbereichen *Grafik & Design* und *Kommunikation & Koordination* zugeordnet. Schaut man sich jedoch die Wissensbestandskarte in Abbildung 4 an, wird ersichtlich, dass Andreas Herder im Bereich *Inhalt* ebenfalls über sehr gute und auch im Bereich der *Animation* über gute Kenntnisse verfügt. Analog verhält es sich mit den anderen Teammitgliedern. Eine Wissensbestandskarte zeigt also viel detaillierter Informationen über Wissensträger und deren Kompetenzen an, als eine Wissensträgerkarte wie sie in Abbildung 3 dargestellt ist.

Ein weiterer Vorteil der Wissensbestandskarte ist, dass mit ihrer Hilfe auch Wissenslücken aufgedeckt werden und so bei Bedarf an deren Schließung gearbeitet werden kann.

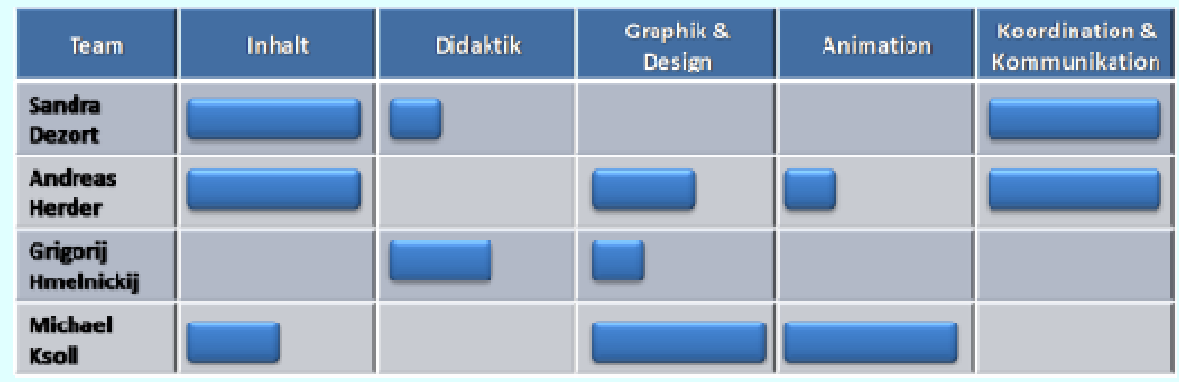

Abbildung 4: Beispiel einer Wissensbestandskarte [KS AG 2011]

Wissensanwendungskarten

Wissensanwendungskarten geben ebenfalls Auskunft über Wissensträger und Wissensressourcen. Der Unterschied zu den zwei bereits vorgestellten Typen von Wissenslandkarten besteht jedoch darin, dass sich die Darstellung in den Wissensanwendungskarten nur auf ein spezifisches Projekt beziehungsweise einen speziellen Projektschritt beschränkt. Auf diese Weise wird Wissen direkt mit Geschäftsprozessen in Verbindung gesetzt. [Ott 2003] Die Reihenfolge der Prozess- bzw. Projektschritte wird in der Wissensanwendungskarte dargestellt und es ist zu jedem Zeitpunkt Aufschluss über die verschiedenen dazugehörigen Wissensträger und Wissensressourcen gegeben.

In Abbildung 5 ist ein Beispiel für eine solche Wissensanwendungskarte dargestellt. Die verschiedenen Prozessschritte des *WBT Erstellungsprozesses* werden in einem Fluss-

diagramm dargestellt und jeder einzelne Schritt steht in einer Beziehung mit seinen Wissensträgern.

Eine solche Form der Wissenslandkarte hat einen hohen praktischen Nutzen und kann dazu beitragen Geschäftsprozesse effizienter zu gestalten.

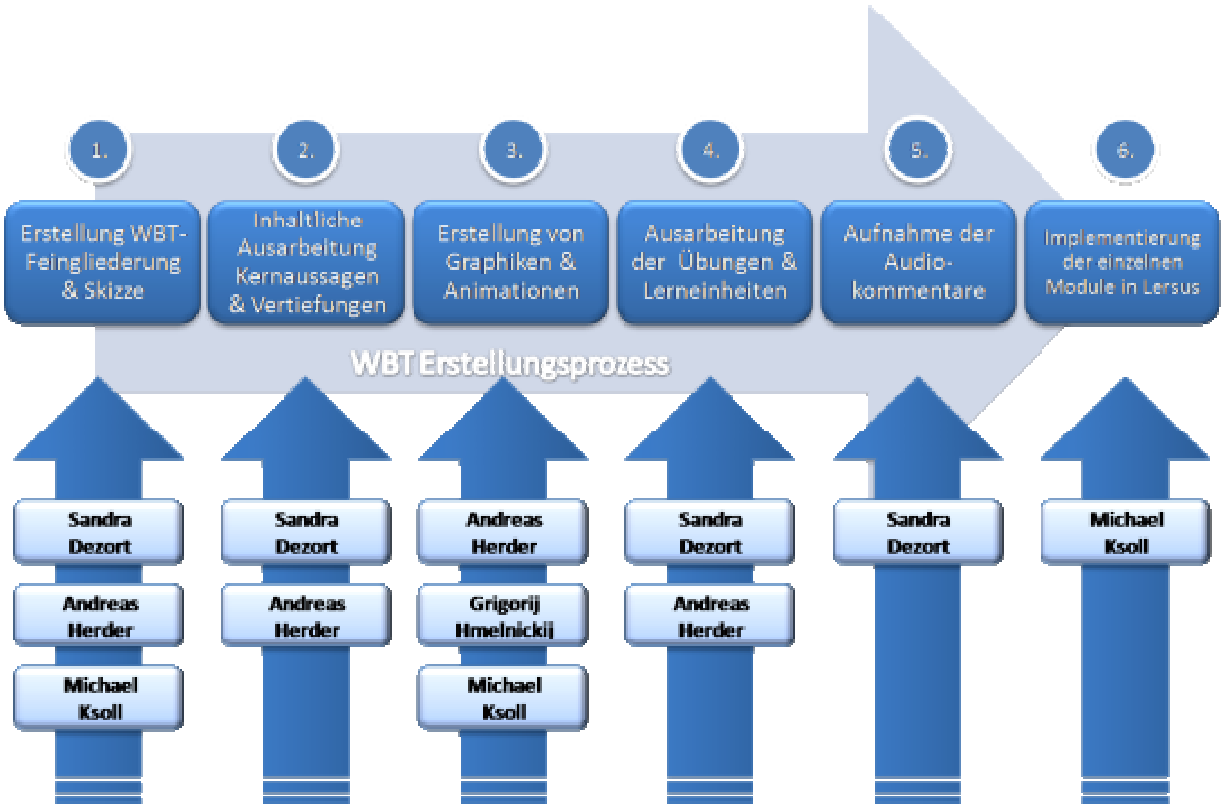

Abbildung 5: Beispiel einer Wissensanwendungskarte [KS AG 2011]

Wissensstrukturkarten

Mit Hilfe von Wissensstrukturkarten können Beziehungen, Abhängigkeiten und komplexe Zusammenhänge zwischen Sachverhalten aufgedeckt werden. „Es sollen nicht mehr nur Elemente isoliert voneinander betrachtet werden, sondern es soll strukturelles Wissen in seiner vollen Komplexität erfasst und bewertet werden.“ [Ott 2003] Ziel ist es verschiedene Aufgabenfelder darzustellen und deren Inhalte und strukturelle Zusammenhänge transparent zu machen. [Ott 2003]

Die Abbildung 6 zeigt eine solche Wissensstrukturkarte. Es wird deutlich, dass hier die Beziehungsnetze zwischen den einzelnen Strukturelementen im Mittelpunkt stehen. Diese Karten geben „Antwort auf die Frage *Welche Wissensgebiete gibt es und wie sind sie strukturiert?*“ [KS AG 2011].

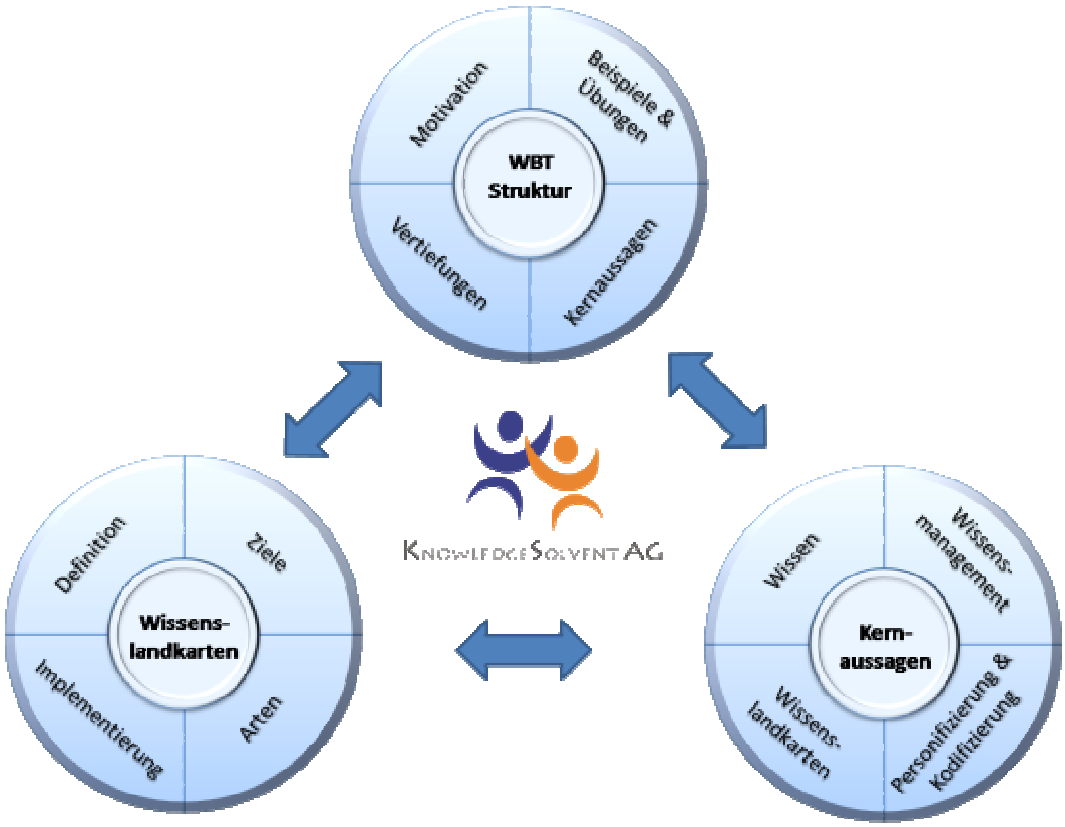

Abbildung 6: Beispiel einer Wissensstrukturkarte [KS AG 2011]

Wissensentwicklungskarten

Wissensentwicklungskarten haben den Aufbau, die Erweiterung und die Weiterentwicklung von Wissen zum Ziel. Die Wissensgebiete werden in verschiedene Teilgebiete unterteilt, die Karte stellt eine Art Wegweiser durch diese dar. Aus Wissensentwicklungskarten wird ersichtlich, „welches Wissen bereits vorhanden ist und welches noch entwickelt werden muss“ [Ott 2003]. Somit können Kompetenzen erweitert und Wissenslücken geschlossen werden. Ein Beispiel für eine solche Wissensentwicklungskarte ist in Abbildung 7 dargestellt.

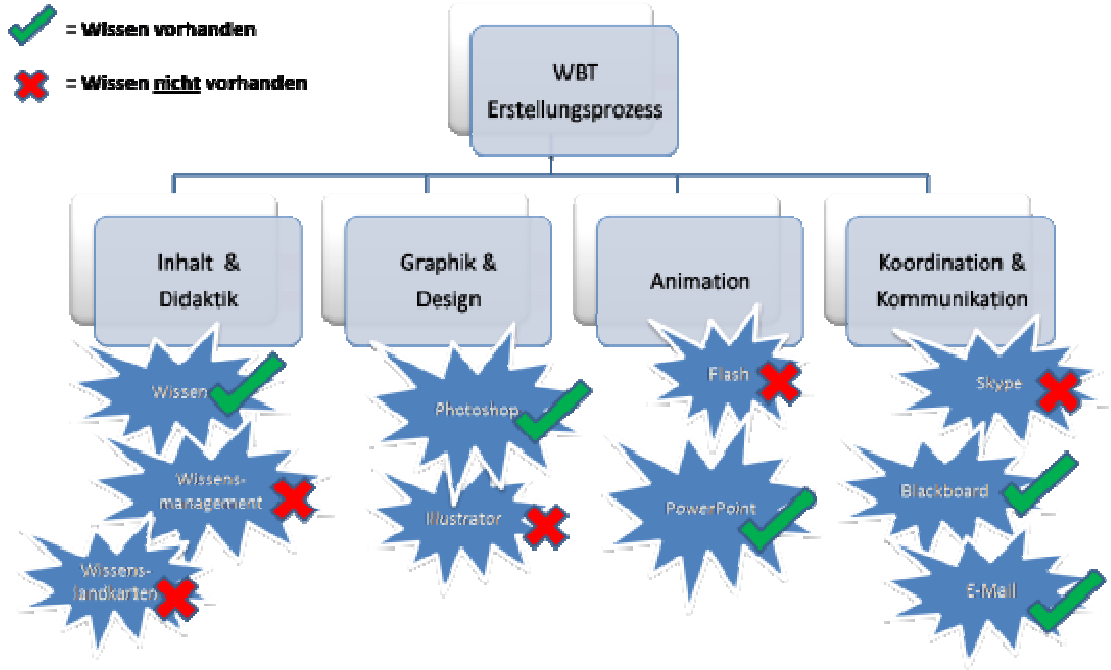

Abbildung 7: Beispiel einer Wissensentwicklungskarte [KS AG 2011]

3.5 Anwendungsbeispiel KnowTech 2004

3.5.1 Erstellen der Wissenslandkarte KnowTech 2004

Im Folgenden möchte ich die Schritte der automatische Textanalyse noch einmal kurz zusammenfassen und konkret anhand des Beispiels *Eine Wissenslandkarte der KnowTech 2004* [Biemann et al. 2004] darlegen. Bei der KnowTech 2004 handelt es sich um eine Tagung mit dem Themenschwerpunkt Wissensmanagement. Es wird aufgezeigt, wie sich „die wesentlichen Themenschwerpunkte und deren Schlüsselbegriffe, die Zuordnung von Autoren zu Themenschwerpunkten, die Zuordnung von Autoren zu Firmen sowie die Vernetzung von Themen“ [Biemann et al. 2004] grafisch in einer Wissenslandkarte darstellen lassen. Dabei wird im hier vorgestellten Ansatz die „Datenbasis durch die Anwendung von Text Mining auf bereitgestellten Textressourcen selbst geschaffen.“ [Heyer et al. 2006]

Bevor es mit der Erstellung des Grafen beginnen kann, muss zunächst wie in Kapitel 3.2.1 beschrieben eine textuelle Vorverarbeitung erfolgen. Die Vorverarbeitung muss für alle Beiträge durchgeführt werden und umfasst folgende Schritte:

- Konvertierung der Texte in ein ASCII-Format
- Extraktion von Titel, Autoren und deren zugehörige Organisationen
- Segmentierung der Texte in Sätze und Wortformen (vgl. Segmentierung von Text in Sätze und Wortformen (S.13))
- Grundformreduktion (vgl. Grundformreduktion (S.17))

Der nächste Schritt besteht aus der Beschlagwortung der Texte. Dies geschieht mit Hilfe der Differenzanalyse (vgl. Differenzanalyse (S.18)). Als Referenzkorpus dient in diesem Beispiel der Deutsche Wortschatz aus dem Projekt *Deutscher Wortschatz*.

Aufgrund der Tatsache, dass es sich bei wichtigen Fachausdrücken häufig um Komposita[1] handelt, wird im nächsten Schritt eine Kompositazerlegung durchgeführt. Die Frequenzen der einzelnen Bestandteile dieser Komposita werden erfasst und diejenigen Komposita, die hochfrequente Teile enthalten, werden im Anschluss ebenfalls als Schlagwörter extrahiert.

Für die Darstellung der Ergebnisse als Graf, wird nun jeder Beitrag mit seinen Autoren und seinen Schlagwörtern verbunden. Die Autoren werden dann mit ihren Organisationen und Co-Autoren verbunden. Als Ergebnis hat man eine grafische Darstellung, anhand derer charakteristische Schlagwörter jedes Beitrags zu erkennen sind, sowie der Autor eines jeden Textes und die Zugehörigkeit dieses Autors zu einer Organisation.

Fachtagungen und Messen drehen sich oft um ein zentrales Thema. Infolgedessen werden einige Schlagwörter sicher aus nahezu allen Beiträgen extrahiert und es existiert eine starke Vernetzung zwischen fast allen Beiträgen. Diese häufigen Schlagwörter nennt man *Hubs* und die Folge sind oft sehr dichte und somit unübersichtliche Grafen. Da die Hubs das Themengebiet jedoch nur sehr allgemein beschreiben und keinen Strukturierungswert besitzen, lautet die Lösung diese Hubs aus dem Grafen zu löschen. Häufig reicht es, die fünf größten Hubs zu entfernen. Am Beispiel der KnowTech 2004 wären das absteigend nach ihrer Verbindungsstärke geordnet: *Wissensmanagement, Knowledge, Communities, Daten und Benutzer*.

[1] Komposita: zusammengesetzte Substantive

Bei dem Tagungsthema „Wissensmanagement“, ist es nicht überraschend, dass genau dies auch der größte Hub ist. Diese Hubs dienen keiner konkreten Beschreibung des Themas.

Schon aufschlussreicher hingegen sind Hubs bei denen man sich entschied sie im Grafen beizubehalten. Dabei handelt es sich beispielsweise um: *wissensintensiv, Suche, semantisch, Geschäftsprozess* und *Informationsraum*.

Nach der Durchführung all dieser Arbeitsschritte werden abschließend auch Assoziationen zwischen den Schlagwörtern hinzugenommen. „Zwei Schlagwörter erhalten dann eine assoziative Verbindung im Grafen, wenn sie gemeinsame Schlagwörter eines Beitrages sind und signifikant häufig miteinander in Sätzen der Kollektion vorkommen.“ [Heyer et al. 2006]

In diesem Beispiel entschied man sich die Knoten in folgende Klassen zu unterscheiden:

- Beitragstitel
- Person
- Organisation
- Schlagwort

Die Relationen, die zwischen diesen Konten existieren können lauten:

- Schlagwort-von
- Autor-von
- Arbeitet-für
- Co-Autor
- Ungetypte Assoziation

Der entstandene Graf wird in das Programm SemanticTalk (vgl. Kapitel SemanticTalk (S.35)) geladen.

3.5.2 Visualisierung der Wissenslandkarte KnowTech 2004

Die Abbildung 8 zeigt eine Gesamtübersicht der entstandenen Wissenslandkarte der KnowTech 2004 mit ausgewählten markierten Abschnitten.

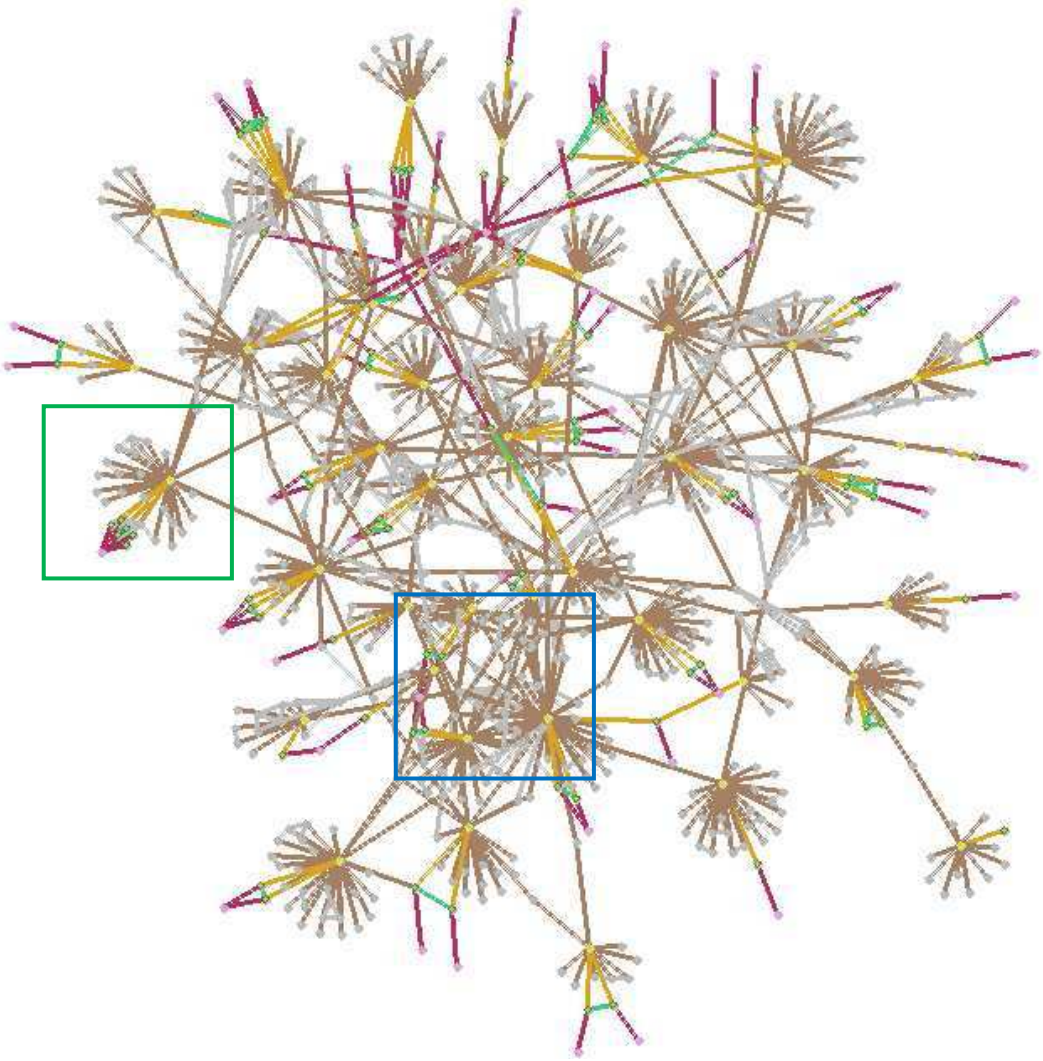

Abbildung 8: Gesamtüberblick Wissenslandkarte KnowTech 2004 [Biemann et al. 2004]

Deutlich zu erkennen sind die entstandenen sternförmigen Gebilde, deren Mitten aus den Beitragstiteln bestehen.

Je mehr gemeinsame Schlagwörter ein Beitrag mit anderen Beiträgen gemeinsam hat, desto zentraler ist seine Position im Grafen.

Abbildung 9 zeigt den Ausschnitt, der in Abbildung 8 mit einem grünen Viereck markiert wurde. Es handelt sich hierbei um den Einzelbeitrag *Fit für den Wissenswettbewerb*.

In dieser Abbildung ist zu erkennen, dass Beitragstitel gelb markiert wurden, Schlagwörter grau, Autoren grün und Organisationen violett. Dieses Farbschema gilt ebenfalls für die Abbildungen 8, 10 und 11.

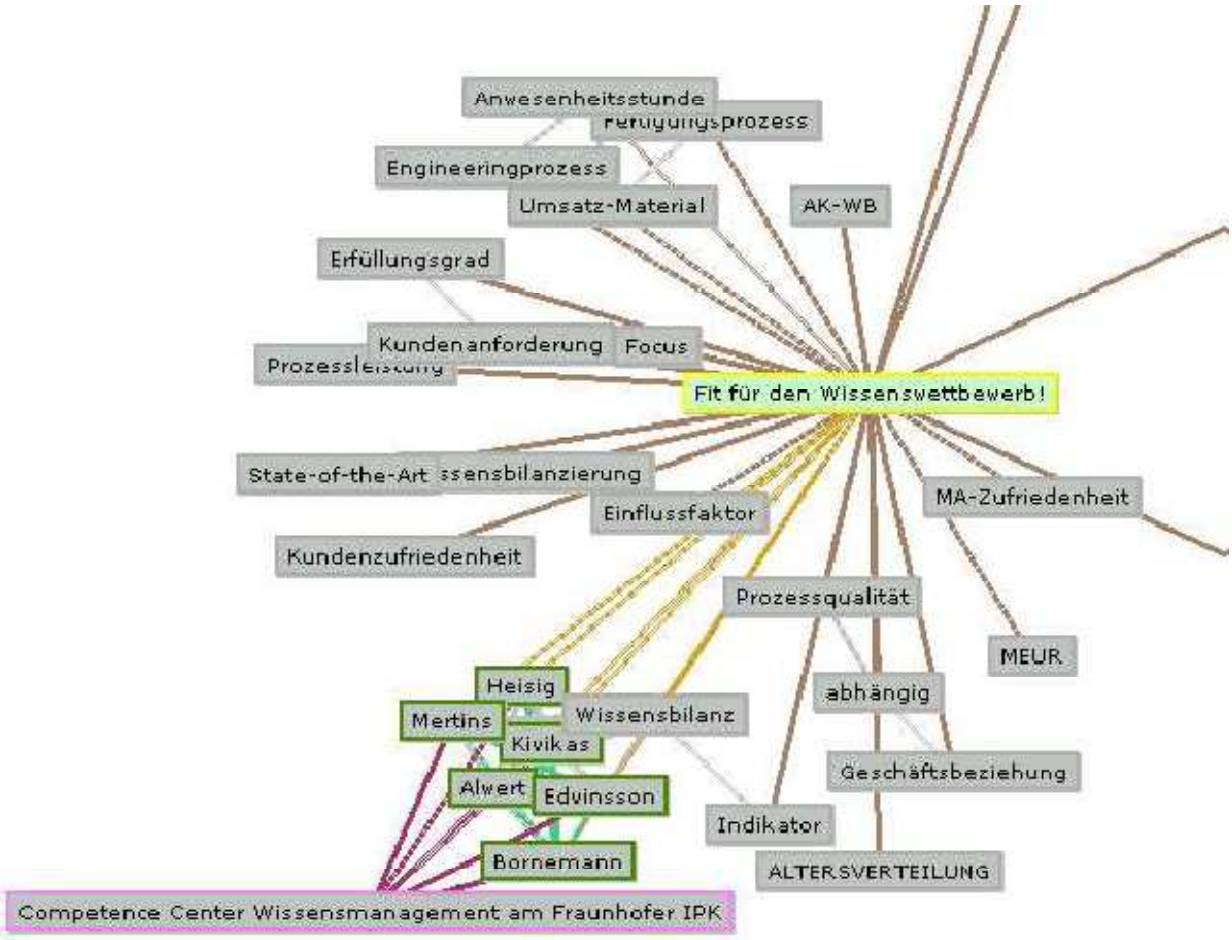

Abbildung 9: Ausschnitt aus der Wissenslandkarte KnowTech 2004 [Biemann et al. 2004]

Anhand dieser Abbildung ist deutlich zu erkennen, dass der Beitragstitel *Fit für den Wissenswettbewerb* in der Mitte des sternförmigen Objekts steht. Die Schlagwörter, Autoren und Organisationen verteilen sich nun wie folgt um diesen Beitragstitel herum:

Alle Autoren sind gemeinsam in einer Ecke gruppiert und mit ihrer Organisation, sowie mit dem zentral stehenden Beitragstitel verbunden. Die Schlagwörter verteilen sich um den Beitragstitel herum, wobei assoziierte Schlagwörter nah beieinander stehen. So zum Beispiel die Gruppe: *Anwesenheitsgruppe, Fertigungsprozess, Energieprozess, Umsatz-Material.*

Abbildung 10 und 11 stellen den in der Abbildung 8 mit einem blauen Viereck markierten Bereich dar.

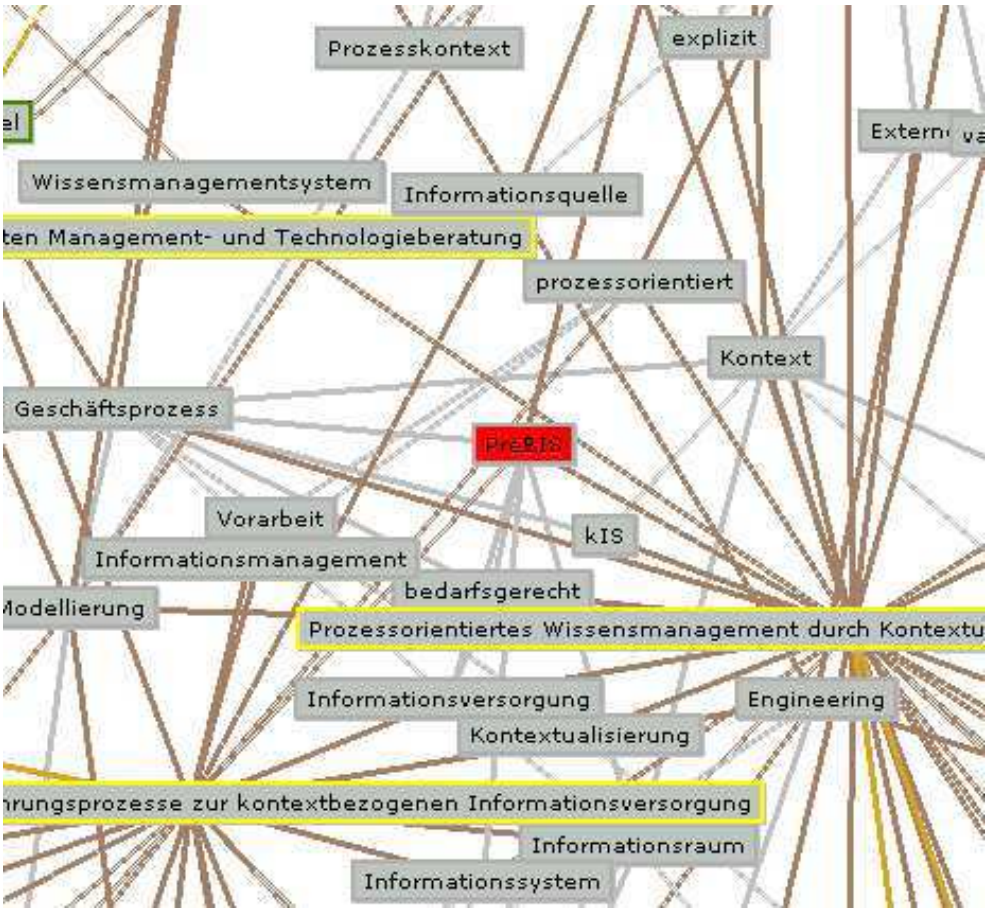

Abbildung 10: Umfeld des Wortes PreBIS mit allen Stichwörtern [Biemann et al. 2004]

Dem Forschungsprojekt PreBIS ist es gelungen mehrere Beiträge bei der KnowTech 2004 zu verorten. Anhand dieses Projekts kann man deutlich erkennen, dass Beitragstitel, die sich mit einem ähnlichen Thema beschäftigen im Graf nah beieinander stehen. Abbildung 10 zeigt das Umfeld von PreBIS mit den Beitragstiteln und allen Schlagwörtern, wohingegen in Abbildung 11 die Beitragstitel, Autoren und Organisationen veranschaulicht werden.

Aus der Abbildung 10 „ist ersichtlich, dass sich das PreBIS-Projekt mit der kontextabhängigen und prozessorientierten Modellierung von Informationen befasst“ [Heyer et al. 2006].

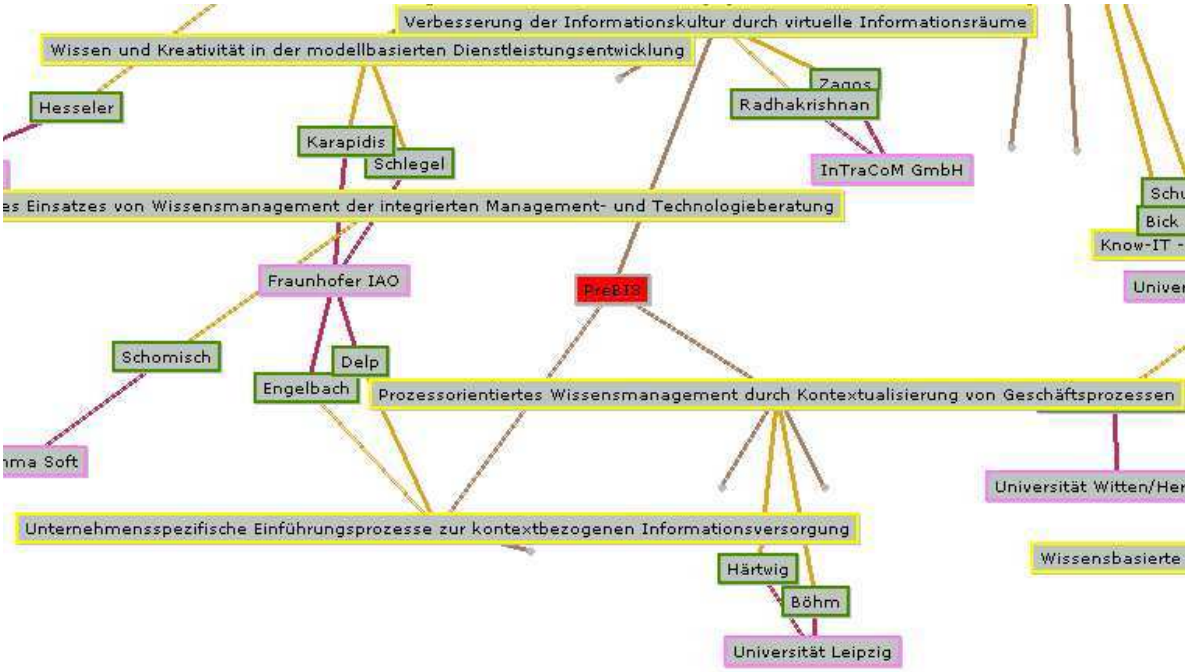

Abbildung 11: Umfeld des Wortes PreBIS mit Beitragstiteln, Autoren, Organisationen
[Biemann et al. 2004]

Diese Abbildung zeigt, dass sich drei Organisationen im näheren Umfeld von PreBIs aufhalten: das *Frauenhofer IAO*, die *Universität Leipzig* und die *InTraCom GmbH*. Der Grund hierfür ist, dass das Wort PreBIS in allen drei Beiträgen dieser Institutionen korrekt extrahiert wurde und diese auch sonst noch viele weitere gemeinsame Schlagwörter haben. So zum Beispiel *Kontextualisierung* und *Informationsversorgung*.

Existieren im Gegensatz dazu nur schwache Assoziationen zwischen einzelnen Knoten, werden diese Knoten im Graf nicht nah beieinander positioniert.

3.5.3 Anwendung der Wissenslandkarte KnowTech 2004

Die Erstellung dieser Wissenslandkarte ist sowohl für die Konferenzteilnehmer als auch für die Besucher von großem Nutzen. Möchte sich ein Konferenzteilnehmer über Beiträge informieren, die seinem eigenen Vortragsthema ähneln, wird er diese in Anbetracht dessen, dass inhaltlich ähnliche Beiträge im Graf nah beieinander stehen, ganz in der Nähe seines eigenen Beitrages auf der Wissenslandkarte finden.

Aber auch Besucher können sich aufgrund dieser semantischen Distanz im Graf einen raschen Überblick darüber verschaffen, welche Beiträge für sie interessant sind.

3.6 Softwaretools

In diesem Kapitel meiner Arbeit möchte ich auf einige Softwaretools hinweisen, die die Erstellung solcher Wissenslandkarten unterstützen können.

3.6.1 TextToOnto

Mit Hilfe der Software TextToOnto, welche als kostenfreie Open-Source Anwendung zur Verfügung steht [TextToOnto 2011], lassen sich Ontologien anhand von natürlichsprachigen Korpora halb- und vollautomatischen erstellen.

An dieser Stelle möchte ich den Begriff *Ontologie* näher erläutern. Die Definition des Wortes Ontologie lautet nach [Witte, Mülle 2006] wie folgt: „Um miteinander zu kommunizieren und Wissen auszutauschen ist zwischen den Kommunikationspartnern ein gewisses Einverständnis über die Grundstrukturen der Welt erforderlich. Formal repräsentiertes Wissen benötigt – wenn es austauschbar sein soll – als Grundlage eine abstrakte, vereinfachte Sicht auf einen geeigneten Ausschnitt der Welt. Eine Ontologie ist eine explizite Formalisierung einer solchen Sicht, eine formale Beschreibung der grundlegenden existierenden Konzepte und ihrer Beziehungen.“ Ontologien werden in generische Ontologien und Domänenontologien unterschieden. Dabei stellen generische Ontologien Beziehungen allgemeiner Begriffe dar und finden in vielen verschiedenen Fachgebieten Anwendung. Dahingegen beschränken sich Domänenontologien auf ein bestimmtes Fachgebiet, wie beispielsweise Medizin oder Politik.

Zur Veranschaulichung einer Ontologie soll die Abbildung 12 dienen. In der sind drei typische Elemente von Ontologien zu erkennen, die laut [Witte, Mülle 2006] folgendermaßen erläutert werden:

Konzepte: „Konzepte sind wichtiger Bestandteil aller Ontologien und stellen abstrakte Objekte der modernen Welt dar.“ In der Grafik werden sie durch Vierecke dargestellt (beispielsweise *Politiker* oder *Ministerpräsident*).

Beziehungen: „Benannte oder unbenannte Beziehungen (im Bild die Kanten des Graphen) zwischen diesen Konzepten stellen die Zusammenhänge in dieser Wissensmodellierung dar. Unbenannte Beziehungen stehen in diesem Beispiel für so genannte taxonomische, d.h. Unterkonzeptsbeziehungen.“

Instanzen: „In der Abbildung in abgerundeten Vierecken dargestellt, stellen Instanzen das konkrete Auftreten eines Konzeptes dar. Der Abstraktionsgrad, d.h. die Unterscheidung zwischen Unterkonzepten und Instanzen ist anwendungsabhängig wählbar: So könnte in einem anderen Kontext durchaus Mensch eine Instanz, anstelle eines Unterkonzeptes, von Lebewesen sein."

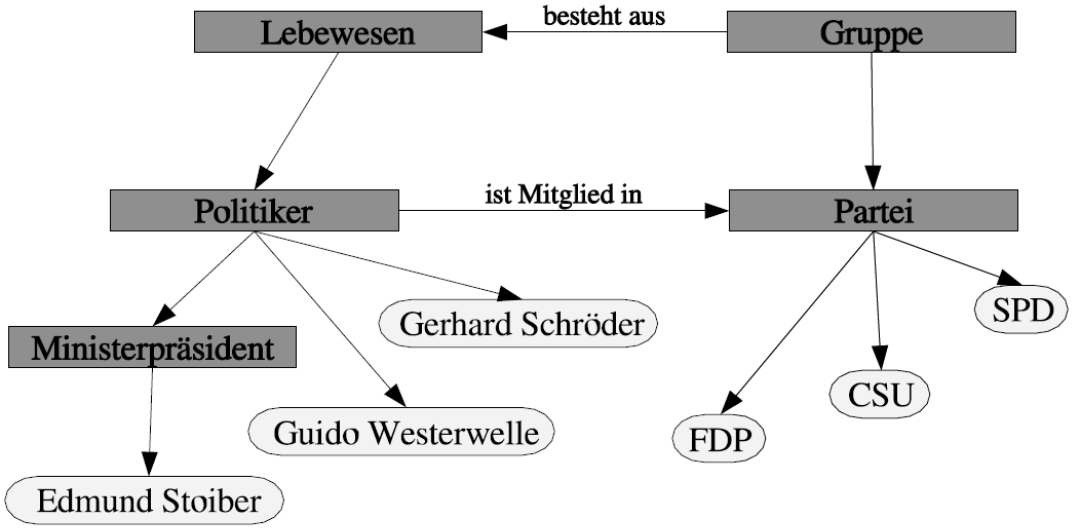

Abbildung 12: Beispielontologie aus der Politik-Domäne [Witte, Mülle 2006]

Nachdem diese Begriffe erläutert wurden, kann mit Funktionsweise des Softwaretools TextToOnto fortgefahren werden.

TextToOnto stellt die Werkzeuge *Term Extraction*, *Instance Extraction*, *Relation Extraction* und *Relation Learning* bereit. Mit diesen Werkzeugen kann nach Eingabe eines Korpus eine Ontologie erstellt werden.

Term Extraction: Mit Hilfe dieses Tools lassen sich Konzepte zu einer Ontologie hinzufügen. Treten Terme mit einer höheren Frequenz als der vom Nutzer festgelegten Häufigkeitsschwelle auf, werden diese in einer Liste angezeigt. Anschließend kann der Nutzer entscheiden, welche Terme als Konzepte dienen sollen.

Instance Extraction: Mit Hilfe dieses Werkzeuges können im Anschluss an das Hinzufügen von Konzepten Instanzen dieser Konzepte identifiziert werden. Dieser Schritt erfolgt vollautomatisch.

Relation Extraction: Mittels der Relation Extraction lassen sich allgemeine, konzeptionelle Beziehungen bestimmen. Dieser Schritt funktioniert jedoch nur semiautomatisch.

Relation Learning: In diesem Schritt werden für Konzepte die oft als Objekte des gleichen Verbes auftauchen gemeinsame Oberkonzepte in der Ontologie angelegt. So könnte beispielsweise für das Verb write das Oberkonzept *writeable* erstellt werden.

[Witte, Mülle 2006]

3.6.2 Protégé

Protégé ist ein Tool zur halbautomatischen Erstellung von Wissensstrukturen (Ontologien). „Die Ontologien werden aus einer Menge von hierarchisch angeordneten Konzepten (in Protégé: *Classes*), ihnen zugeordneten Instanzen (*Instances*) und zwischen diesen Instanzen bestehenden Beziehungen, die deren Eigenschaften ausdrücken (*Slots* bzw. *Properties*), aufgebaut. Für jedes dieser Elemente gibt es im Editor eine Registerkarte (einen Tab) unter der sie beschrieben werden können.“ [Ahrens 2011]

Protégé steht als freie Software zur Verfügung [Protégé 2011] und gehört mittlerweile zu den meist genutzten Ontologieeditoren.

3.6.3 SemanticTalk

Das Softwaretool SemanticTalk stellt die Grafenstruktur von Dokumentensammlungen auf einer zweidimensionalen Fläche dar.

Zunächst muss jedoch analog zu Kapitel 3.5.1 Erstellen der Wissenslandkarte KnowTech 2004 ein Graf erstellt werden, sodass der Inhalt der Dokumente grafisch-visuell dargestellt ist. Dieser Graf ermöglicht einen raschen Überblick über den Inhalt der Texte und kann anschließend in SemanticTalk geladen werden. Dort wird er automatisch positioniert.

Mit Hilfe der Software SemanticTalk können den Wörtern und Assoziationen im Graf nun Typen zugeordnet werden. Im Fall der Wissenslandkarte KnowTech 2004 lauteten die Typen für die Knoten *Beitragstitel, Person, Organisation oder Schlagwort* und für die As-

soziationen *Schlagwort-von, Autor-von, Arbeitet-für oder Co-Autor.* Anhand dieser Typen können später Knoten und/oder Assoziationen ausgeblendet werden, sodass Ansichten auf Substrukturen ermöglicht werden. Mit Hilfe solcher Substrukturen können dann beispielsweise Personennetzwerke oder Ressourcenmodelle erstellt werden.

Eine weitere Funktion dieser Software ist, dass sowohl Wörter als auch Assoziationen des Grafen mit Informationen hinterlegt werden können. „Auf diese Weise kann SemanticTalk als visueller Browser auf einer Dokumentenkollektion eingesetzt werden." [Biemann et al. 2004]

SemanticTalk ist eine kostenpflichtige Software, eine Testversion des Programms wird jedoch als kostenfreie Open-Source Anwendung bereit gestellt. [SemTalk 2011]

3.7 Vor- und Nachteile von Wissenslandkarten

Neben den Vorteilen die eine strukturierte Darstellung großer Datenmengen in Form einer Wissenslandkarten haben, sollten jedoch auch die Kritikpunkte nicht vernachlässigt werden. An dieser Stelle möchte ich die Vor- und Nachteile der Wissenslandkarten noch einmal zusammenfassen.

Vorteile

- Wissenslandkarten sind zeit- und raumunabhängig aufzurufen und lassen sich im Gegensatz zu Informationen auf Papier besser und schneller aktualisieren.
- Sie bieten einen schnellen Wissensaustausch und raschen Zugriff auf benötigte Informationen.
- Wissenslandkarten beschleunigen den Prozess vom impliziten zum expliziten Wissen.
- KnowledgeMaps stellen eine gute Möglichkeit der Zusammenfassung mehrerer Beiträge zu einem Thema dar. [Wissenslandkarte 2008]
- Wissensressourcen können effizienter genutzt werden.
- Aber auch Wissensdefizite können mit Hilfe von Wissenslandkarten identifiziert werden.
- Experten spezifischer Fachgebiete lassen sich ausfindig machen.
- Wissenslandkarten schaffen einen gemeinsamen Kontext.
- Große Datenmengen und komplexe Zusammenhänge können grafisch-visuell schneller und exakter erfasst werden, als es verbal oder durch Zahlenwerte möglich ist [Wissenslandkarte 2008].

Nachteile

- Qualitativ hochwertige Wissenslandkarten erfordern einen großen Aufwand; ihre Erstellung ist schwierig, anspruchsvoll und zeitintensiv (Erfassung von Wissensbeständen, Messung von Mitarbeiterfähigkeiten, etc.)
- Wissen wächst stetig an, somit ist die Erstellung von Wissenslandkarten kein einmaliger Prozess und ihre Aktualisierung stellt eine Herausforderung dar.
- Die Darstellung dynamischer Prozesse gestaltet sich schwierig.
- Es besteht die Gefahr der Fehlinterpretation der Wissenslandkarte.
- Es kann leicht zu einem Informationsüberfluss kommen, so dass die Karten unübersichtlich werden.

4 Zusammenfassung

Im ersten Teil meiner Arbeit wurde anhand von 30 biologischen Texten mit dem Programm antconc 3.2.1w eine Textanalyse durchgeführt. So gelang es aus diesen Texten wichtige Informationen und Zusammenhänge zu extrahieren.

Der zweite Teil meiner Arbeit beschäftigt sich hauptsächlich damit, darzustellen, wie anhand von Wissenslandkarten solche Informationen übersichtlich und strukturiert grafisch dargestellt werden können.

Sollen qualitativ hochwertige Wissenslandkarten erstellt werden, kann dies mehrere Experten über einen Zeitraum mehrerer Wochen beanspruchen. Obwohl dieser wesentliche Zeit- und Kostenfaktor auf keinen Fall vernachlässigt werden darf, lässt sich festhalten, dass Wissenslandkarten ein breitgefächertes Anwendungsgebiet haben und ein nützliches und wertvolles Instrument des Wissensmanagements darstellen. Die uns zur Verfügung stehenden Daten und somit auch das Wissen steigen stetig an, so dass es immer schwieriger wird einen Überblick darüber zu behalten. Mit Hilfe von Wissenslandkarten wird eine Transparenz über bestehendes Wissen geschaffen, so dass dieses Wissen von den Nutzern effizienter genutzt werden kann. Desweiteren zielen sie auf die Fähigkeit unseres Gehirns ab große Datenmengen und komplexe Zusammenhänge rascher erfassen zu können, wenn diese grafisch-visuell dargestellt sind.

Literatur

[Ahrens 2011] Ahrens M: Semi-automatische Generierung einer OWL-Ontologie auus domänenspezifischen Texten am Beispiel von HUMINT-Meldungen. URL:< http://storage.sk.uni-bonn.de/abschlussarbeiten/magisterarbeit_ahrens.pdf>, verfügbar am 20.08.2011, 10.00

[AntConc] Using AntConc (Version 3.2.1w) step by step. URL:<http://fss.plone.uni-giessen.de/fss/faculties/f05/engl/ling/help/materials/restricted/antconc/file/usingAntConc.pdf>, verfügbar am 20.08.2011, 10.00

[Anthony 2011] Anthony L: AntConc Homepage. URL:<http://www.antlab.sci.waseda.ac.jp/antconc_index.html>, verfügbar am 20.08.2011, 10.00

[Biemann et al. 2004] C.Biemann; G.Heyer; F.Schmidt; H.Friedrich Witschel: Eine Wissenslandkarte der KnowTech. URL:< http:// wortschatz.uni-leipzig.de/~fwitschel/papers/ KnowtechLandkarte.pdf>, verfügbar am 20.08.2011, 10.00

[Hackl 2006] Hackel, E: Die konzeptionelle Entwicklung einer Wissenslandkarte zur Unterstützung der Wiederverwendung von Projektergebnissen. URL:< http://know-center.tugraz.at/wp-content/uploads/2010/12/Diplomarbeit-Elisabeth-Hackl.pdf, verfügbar am 20.08.2011, 10.00

[Heyer et al. 2006] Heyer; Quasthoff; Witting: Text Mining: Wissensrohstoff Text. – 1. Aufl. – Herdecke: W3L, 2006

[Jansen, Smith 2008] Jansen; Smith: Biomedizinische Ontologie: Wissen strukturieren für den Informatik-Einsatz. – Zürich: vdf Hochschulverlag AG, 2008

[Kinner, Haag] Kinner; Haag: Knowledge Mapping – Knowledge Maps (Wissenskarten). URL:< http://v.hdm-stuttgart.de/seminare/wm/ws9900/knowledgemapping.html>, verfügbar am 20.08.2011, 10.00

[KS AG] KnowledgeSolvent AG: Wissenslandkarten. URL:< http://web-imtm.iaw.rub.de/fmdb/wbt/html/content/content8.html>, verfügbar am 20.08.2011, 10.00

[Markowski] Markowski: Wegweiser für Ihre Potenziale. URL:< http://www.robert-markowski.net/wissenslandkarten.html>, verfügbar am 20.08.2011, 10.00

[Navab et al. 2011] Navab R; Strumpf D; Bandarchi B; Zhu CQ; Pintilie M; Ramnarine VR; Ibrahimov E, Radulovich N; Leung L; Barczyk M; Panchal D; To C; Yun JJ; Der S; Shepherd FA; Jurisica I; Tsao MS: Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. URL:<http://www.ncbi.nlm.nih.gov/pubmed?term=Prognostic%20gene-expression%20signature%20of%20carcinoma-associated%20fibroblasts>, verfügbar am 20.08.2011, 10.00

[NCBI 2011] NCBI: PubMed. URL:< http://www.ncbi.nlm.nih.gov/pubmed/>, verfügbar am 20.08.2011, 10.00

[Ott 2003] Ott F: Wissenslandkarten als Instrument des kollektiven Wissensmanagements. URL:< http://fhib5jg.factlink.net/fsDownload/DA_Wissenslandkarten.pdf?forumid=286&v=1&id=166113>, verfügbar am 20.08.2011, 10.00

[Preissler, Roehl, Seemann 1997] Preissler; Roehl; Seemann: Wissenslandkarte. URL:<http://www.enbiz.de/wmk/papers/public/HakenHelmSeil/hakenhelmseil.5.html, verfügbar am 20.08.2011, 10.00

[Protégé 2011] Protégé: Protege 3.4.7. URL:< http://protege.stanford.edu/download/protege/3.4/installanywhere/Web_Installers/>, verfügbar am 20.08.2011, 10.00

[SemTalk 2011] SemTalk: Sem Talk-Unsere Testversion. URL:< http://www.semtalk.de/demoversion.html>, verfügbar am 20.08.2011, 10.00

[Stampf 2006] Stampf, I: Erstellung, Implementierung und Anwendung der Wissenslandkarte „Big Picture“. URL:< http://bibliothek.fh-burgenland.at/fileadmin/Download/bibliothek/diplomarbeiten/AC05370270.pdf, verfügbar am 20.08.2011, 10.00

[TextToOnto 2011] KAON Tool Suite: KAON. URL:< http://kaon.semanticweb.org>, verfügbar am 20.08.2011, 10.00

[Universität Duisburg-Essen 2008] Universität Duisburg-Essen: Intelligente Recherchestrategien für e-Humanities. URL:<http://duepublico.uni-duisburg-essen.de/servlets/derivateServlet/Derivate-19651/KUWALU Website/pdf/antconc-test.pdf> verfügbar am 20.08.2011, 10.00

[Wissenslandkarte 2008] Wissenslandkarte. URL:< http://www.brunnbauer.ch/wissensmanagement/index.php?title=Wissenslandkarte>, verfügbar am 20.08.2011, 10.00

[Witte, Mülle 2006] Witte, René: Mülle Jutta: Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten. URL: <http://digbib.ubka.uni-karlsruhe.de/volltexte/1000005161>, verfügbar am 20.08.2011, 10.00

Anlagen

Anlagen, Teil 1

Bei dem hier aufgeführten Abstract *Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer* handelt es sich um jenen Abstract, an dem die Ergebnisse der Textanalyse aus Kapitel zwei auszugsweise dargestellt wurden.

1. Proc Natl Acad Sci U S A. 2011 Apr 7. [Epub ahead of print]

Prognostic ***gene-expression*** signature of ***carcinoma***-associated ***fibroblasts*** in non-small cell ***lung cancer***.

Navab R, Strumpf D, Bandarchi B, Zhu CQ, Pintilie M, Ramnarine VR, Ibrahimov E, Radulovich N, Leung L, Barczyk M, Panchal D, To C, Yun JJ, Der S, Shepherd FA, Jurisica I, Tsao MS.

The Campbell Family Institute for ***Cancer Research***, Ontario ***Cancer Institute*** at Princess Margaret Hospital, University Health Network, Toronto, ON, Canada M5G 2M9.

The tumor microenvironment strongly influences ***cancer development***, progression, and metastasis. The role of ***carcinoma***-associated ***fibroblasts*** (CAFs) in these processes and their clinical impact has not been studied systematically in non-small cell ***lung carcinoma*** (NSCLC). We established primary cultures of CAFs and matched normal ***fibroblasts*** (NFs) from 15 resected NSCLC. We demonstrate that CAFs have greater ability than NFs to enhance the tumorigenicity of ***lung cancer*** cell lines. Microarray ***gene-expression*** analysis of the 15 matched CAF and NF cell lines identified 46 differentially ***expressed genes***, encoding for proteins that are significantly enriched for extracellular proteins regulated by the TGF-β signaling pathway. We have identified a subset of 11 genes (13 probe sets) that formed a prognostic ***gene-expression*** signature, which was validated in multiple independent NSCLC microarray datasets. Functional annotation using protein-protein interaction analyses of these and published ***cancer*** stroma-associated ***gene-expression*** changes revealed prominent involvement of the focal adhesion and MAPK signaling pathways. Fourteen (30%) of the 46 ***genes*** *also were differentially* ***expressed*** in laser-capture-microdissected corresponding primary tumor stroma compared with the matched normal lung. Six of these 14 genes could be induced by TGF-β1 in NF. The results establish the prognostic impact of CAF-associated ***gene-expression*** changes in NSCLC patients.
PMID: 21474781 [PubMed - as supplied by publisher] [Navab et al. 2011]

Anlagen, Teil 2

Tabelle 8 beinhaltet die von mir definierten Schlüsselwörter, um die die Ergebnisse der Textanalyse im Anschluss visuell in Form eines Venn-Diagramms darstellen zu können.

Tabelle 8: Vorkommen der Schlüsselwörter

CGE	**cancer**	**gene**	**expression**	**cell**	**breast**	**liver**	**lung**
CGE 1	x	x	x	x			x
CGE 2	x	x		x	x		
CGE 3	x	x	x	x			
CGE 4	x	x	x			x	
CGE 5	x	x	x				
CGE 6	x	x	x	x			
CGE 7	x	x	x	x			x
CGE 8	x	x	x	x			
CGE 9	x	x	x	x	x	x	x
CGE 10	x	x	x	x			
CellGE	**cancer**	**gene**	**expression**	**cell**	**breast**	**liver**	**lung**
CellGE 1		x	x	x			
CellGE 2		x	x	x			
CellGE 3		x	x	x			
CellGE 4		x	x	x			
CellGE 5		x	x	x			
CellGE 6		x	x	x			
CellGE 7	x	x	x	x			x
CellGE 8	x	x	x	x	x		
CellGE 9		x	x	x		x	
CellGE 10	(mutant)	x	x	x			
GEA	**cancer**	**gene**	**expression**	**cell**	**breast**	**liver**	**lung**
GEA 1		x	x				
GEA 2		x	x	x			
GEA 3		x	x	x			
GEA 4		x	x	x			
GEA 5	x	x	x	x			x
GEA 6	x	x	x	x	x	x	x
GEA 7	(tumor)	x	x	x		x	x
GEA 8		x	x				
GEA 9	x	x	x				
GEA 10		x	x				

CGE	regulation	pcr	rna	leukemia	transcription	immuno	vascular
CGE 1							
CGE 2	x						
CGE 3				x			
CGE 4							
CGE 5						x	
CGE 6	x		x				
CGE 7							
CGE 8	x		x		x		
CGE 9					x		
CGE 10	**x**					x	

CellGE	regulation	pcr	rna	leukemia	transcription	immuno	vascular
CellGE 1			x				
CellGE 2		x		x	x	x	
CellGE 3	x		x		x		x
CellGE 4							x
CellGE 5	x		x	x			x
CellGE 6			x				x
CellGE 7							
CellGE 8	x		x				
CellGE 9	x					x	
CellGE 10							

GEA	regulation	pcr	rna	leukemia	transcription	immuno	vascular
GEA 1		x				x	
GEA 2		x		x	x	x	
GEA 3	x	x			x	x	
GEA 4							
GEA 5							
GEA 6					x		
GEA 7						x	
GEA 8					x		
GEA 9	x					x	
GEA 10	x		x		x		x

Danksagung

An dieser Stelle möchte ich mich gern bei Prof. Dr. rer. nat. Dirk Labudde für die fachliche Unterstützung und Betreuung während des Schreibens dieser Bachelorarbeit bedanken.

Desweiteren möchte ich mich bei meiner Familie bedanken, die mir dieses Studium ermöglichen und mich stets unterstützt und motiviert haben.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, den 20.08.2011

Tina Giersch